Utilizing gene co-expression networks for comparative transcriptomic analyses

A Thesis Submitted to the

College of Graduate and Postdoctoral Studies

In Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy

In the Department of Computer Science

University of Saskatchewan

Saskatoon

By

Katie Ovens

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Or

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

The development of high-throughput technologies such as microarray and next-generation RNA sequencing (RNA-seq) has generated numerous transcriptomic data that can be used for comparative transcriptomics studies. Transcriptomes obtained from different species can reveal differentially expressed genes that underlie species-specific traits. It also has the potential to identify genes that have conserved gene expression patterns. However, differential expression alone does not provide information about how the genes relate to each other in terms of gene expression or if groups of genes are correlated in similar ways across species, tissues, etc. This makes gene expression networks, such as co-expression networks, valuable in terms of finding similarities or differences between genes based on their relationships with other genes.

The desired outcome of this research was to develop methods for comparative transcriptomics, specifically for comparing gene co-expression networks (GCNs), either within or between any set of organisms. These networks represent genes as nodes in the network, and pairs of genes may be connected by an edge representing the strength of the relationship between the pairs. We begin with a review of currently utilized techniques available that can be used or adapted to compare gene co-expression networks. We also work to systematically determine the appropriate number of samples needed to construct reproducible gene co-expression networks for comparison purposes. In order to systematically compare these replicate networks, software to visualize the relationship between replicate networks was created to determine when the consistency of the networks begins to plateau and if this is affected by factors such as tissue type and sample size. Finally, we developed a tool called Juxtapose that utilizes gene embedding to functionally interpret the commonalities and differences between a given set of co-expression networks constructed using transcriptome datasets from various organisms.

A set of transcriptome datasets were utilized from publicly available sources as well as from collaborators. GTEx and Gene Expression Omnibus (GEO) RNA-seq datasets were used for the evaluation of the techniques proposed in this research. Skeletal cell datasets of closely related species and more evolutionarily distant organisms were also analyzed to investigate the evolutionary relationships of several skeletal cell types.

We found evidence that data characteristics such as tissue origin, as well as the method used to construct gene co-expression networks, can substantially impact the number of samples required to generate reproducible networks. In particular, if a threshold is used to construct a gene co-expression network for downstream analyses, the number of samples used to construct the networks is an important consideration as many samples may be required to generate networks that have a reproducible edge order when sorted by edge weight. We also demonstrated the capabilities of our proposed method for comparing GCNs, Juxtapose, showing that it is capable of consistently matching up genes in identical networks, and it also reflects the similarity between different networks using cosine distance as a measure of gene similarity. Finally, we applied our proposed method to skeletal cell networks and find evidence of conserved gene relationships within skeletal GCNs from the same species and identify modules of genes with similar embeddings across species

that are enriched for biological processes involved in cartilage and osteoblast development. Furthermore, smaller sub-networks of genes reflect the phylogenetic relationships of the species analyzed using our gene embedding strategy to compare the GCNs.

This research has produced methodologies and tools that can be used for evolutionary studies and generalizable to scenarios other than cross-species comparisons, including co-expression network comparisons across tissues or conditions within the same species.

Acknowledgements

I would like to thank my supervisors, Dr. Ian McQuillan and Dr. Brian Eames, for supporting me for the duration of my Ph.D research at the University of Saskatchewan. They have guided me through the duration of my research, which has dramatically improved my ability to communicate my research to others and given me the tools to succeed in an academic environment. I have been extremely lucky to have supervisors who cared so much about my work.

This research could not have been completed without the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [funding reference numbers 2016-06172, 435655-201, 2019-05977 and CGS-D scholarship].

I would also like to thank the students in the Eames lab for providing me with the skeletal cell data required to complete the final chapter of this thesis. They have also helped me to improve my presentation abilities and taught me a lot about the biological aspects of bioinformatics research. Patsy Gómez-Picos in particular has continuously provided encouragement and was always willing and enthusiastic to assist in any way she could throughout my research projects.

I would also like to thank my committee: Dr. Anthony Kusalik, Dr. FangXiang Wu, and Dr. Franco Vizeacoumar for their feedback on my research. They have all played a role in polishing my research directions and presentation skills. I would also like to thank my external examiner who took the time out of their schedule to participate in my defence.

My labmates in the Bioinformatics lab have also been and will continue to be valuable collaborators. In particular, I would like to thank my research partner, Dr. Farhad Maleki for collaborating on my papers entitled “pineplot: an R package for visualizing symmetric relationships” and “Juxtapose: A Python tool for gene embedding for co-expression network comparison”, and providing advice regarding my analyses. I am also grateful for the experience with collaborating on his gene set analysis research projects throughout my time as a Ph.D. student. I couldn’t ask for a better research environment than what I have experienced during my time at the University of Saskatchewan.

Contents

List of Tables

List of Figures

List of Abbreviations

Alkphos	alkaline phosphatase
BiNA	Biological Network Alignment
BLAST	Basic Local Alignment Search Tool
COMODO	Conserved Modules Across Organisms
DAG	Directed Acyclic Graph
DNA	Deoxyribonucleic Acid
GCN	Gene Co-expression Network
GEO	Gene Expression Omnibus
GO	Gene Ontology
GTEx	Genotype-Tissue Expression
ISA	Iterative Signature Algorithm
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAGNA	Maximizing Accuracy in Global Network Alignment
MSigDB	Molecular Signatures Database
MUNK	Multi-Species Network Kernel
PCC	Pearson's correlation coefficient
PPI	Protein-protein Interaction Network
RNA	Ribonucleic Acid
RNA-seq	RNA Sequencing
Sgot	aspartate aminotransferase
Sgpt	alanine aminotransferase
TMM	trimmed mean of M-values
WGCNA	Weighted Gene Co-expression Network Analysis

Chapter 1

Introduction

1.1 Background

High-throughput techniques such as RNA-seq and DNA microarray technology have allowed for the large-scale identification of genes and transcripts, their expression patterns, and their interactions [10, 12]. The data from such studies can provide valuable information about the functions of individual genes. Differences in gene expression between phenotypes can also help to identify important genes and the relationships between these genes and biological processes [1]. However, analyses such as differential expression analysis can only report differences in the expression of individual genes and does not consider the coordinated activity of genes. Gene expression involves the coordinated activity of groups of genes, which drives various biological processes and functions. Therefore, the change in expression of a single gene does not fully capture the relationship this change in expression has on the activity of other genes.

One means of considering genes as they relate to the activity of other genes is to treat the genes as a graph or network. Each gene is represented as a node in a network, and pairs of genes may be connected by an edge representing the strength of the relationship between the pairs. Networks have been widely applied to study the complex interactions between genes, proteins, and other small molecules. Gene co-expression network analysis, in particular, has been used to extract new information from differentially expressed genes [2]. Gene co-expression networks may contain many genes, and the edges between the nodes of this type of network are usually weighted using correlation.

These co-expression networks are not static. The biological reality is that not only can there be differences in gene expression levels, but there can also be differences in gene interactions across contexts [9]. This means it is also valuable to compare the gene interactions. Groups of genes conserved over large phylogenetic distances can reveal core components of shared processes. At the same time, important morphological differences can reflect adaptations that correlate strongly with changes in the expression pattern of fundamental regulators. For example, past studies have found that *Hox* genes are conserved in terms of their roles in specifying regions of the body plan of various organisms [4]. This is the case even though the sequences of these genes are not highly conserved ($> 70\%$ divergence) among different phyla [7]. Although the role of the genes do not seem to change, the pattern of their expression does change in certain conditions or times of development, which is one reason that organisms can end up with distinct morphologies. This makes analysing gene–gene interactions important to consider when studying organisms from an evolutionary perspective. Furthermore, co-expression analysis is a useful tool for studying these gene–gene interactions.

The standard approach of gene co-expression network analysis, also commonly known as weighted gene co-expression network analysis (WGCNA) [6], most often follows the steps below.

- A similarity measure is calculated for each pair of genes (e.g. a correlation measure, mutual information, etc.)
- The similarity measures are transformed so that the network exhibits scale-free topology
- Hierarchical clustering is performed with respect to functional organization

The resulting networks are utilized to uncover information about potential regulatory pathways and clusters of genes responsible for related biological processes and functions. We can also determine measures as to whether the modular structure is reproducible and preserved in another dataset. Rarely have these methods been applied to make evolutionary inferences. This is due to the direct analysis of network structure, such as using network alignment to match up nodes and edges of the co-expression networks, being a computationally difficult problem.

1.2 Objectives

The main objective of this thesis is to develop new computational methods and tools useful for comparing transcriptomic data with a particular emphasis on co-expression network comparison. Tools that compare gene expression from a network perspective would be valuable for making evolutionary inferences as more and more gene expression data is gathered from various species. Often biological processes and functions are the result of groups of genes acting in concert, and this should be reflected in network topology. The biological principle of "guilt by association" states that genes are likely to share properties such as genetic or physical interactions if they have related functions [9, 10]. Quantifying such complex interactions can help to compare co-expression networks. Similar ideas have been used in other areas, such as linguistics. Ficklin et al. stated, "You shall know a word by the company it keeps" [3], which can also be claimed for genes. This idea has been heavily utilized in the area of natural language processing to generate numerical representations for words. In this thesis, these same techniques are utilized as a means to compare co-expression networks in a way that allows for a measure of similarity to be established between genes based on how they interact with other genes. Furthermore, a technique to compare the reproducibility of co-expression networks was developed to determine at what sample size co-expression networks become stable enough to be comparable. A visualization tool, *pineplot*, was also designed to compare symmetric matrices, which is an underlying data structure that is also useful for the analysis of biological processes, including the analyses performed in this project [8].

The second objective of this thesis is to apply the tools and techniques produced from my first objective to the study of skeletal cell evolution. The most abundant tissues in vertebrate skeletal tissues are bone, immature cartilage, and mature cartilage. It is currently unclear how these tissues

relate to each other in evolutionary history. One hypothesis is that bone evolved from cartilage [5, 11]. However, to our knowledge, a holistic study from a network perspective to identify conserved gene expression patterns/relationships has not been performed. This research provides a means of identifying conserved sub-graphs across the networks active in these tissue types in different species. It also describes methods for comparing these networks at a global scale, as well as defining measures of similarity between genes based on their topology in the networks.

1.3 Organization

The organization of the remaining chapters in this manuscript style thesis is as follows. Chapter 2 contains a review of the methods used to compare co-expression networks for identifying evidence of constraint or adaptation. We cover the common techniques currently used for comparing co-expression networks, exploring their strengths as well as drawbacks. The methodologies presented in Chapters 3, 4, and 5 of this thesis offer strategies for comparing and analyzing multiple co-expression networks in a systematic and quantitative manner, and provide the tools necessary to address our second objective in Chapter 6. Chapters 3 and 4 contain papers that have appeared in the literature, while Chapters 2, 5, and 6 contain content that is being prepared for submission (for publication). Finally, Chapter 7 provides a summary of the research done in this thesis, highlighting their contributions and limitations, as well as discusses the avenues for future research in GCN comparison.

References

[1] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[2] Mohammad Reza Bakhtiarizadeh, Batool Hosseinpour, Maryam Shahhoseini, Arthur Korte, and Peyman Gifani. Weighted gene co-expression network analysis of endometriosis and identification of functional modules associated with its main hallmarks. *Frontiers in Genetics*, 9:453, 2018.

[3] John Rupert Firth. *Papers in Linguistics, 1934-1951*. Oxford University Press, 1958.

[4] David A Garfield and Gregory A Wray. The evolution of gene regulatory interactions. *BioScience*, 60(1):15–23, 2010.

[5] Patsy Gómez-Picos and B Frank Eames. On the evolutionary relationship between chondrocytes and osteoblasts. *Frontiers in Genetics*, 6:297, 2015.

[6] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, 2008.

[7] Derek Lemons and William McGinnis. Genomic evolution of Hox gene clusters. *Science*, 313(5795):1918–1922, 2006.

[8] Katie L. Ovens, Daniel J. Hogan, Farhad Maleki, Ian McQuillan, and Anthony J. Kusalik. Pineplot: An r package for visualizing symmetric relationships. In *Proceedings of the Tenth International Conference on Computational Systems-Biology and Bioinformatics*, CSBio 19, New York, NY, USA, 2019. Association for Computing Machinery.

[9] Elise AR Serin, Harm Nijveen, Henk WM Hilhorst, and Wilco Ligterink. Learning from co-expression networks: possibilities and challenges. *Frontiers in Plant Science*, 7:444, 2016.

[10] Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, page bbw139, 2017.

[11] Darja Obradovic Wagner and Per Aspenberg. Where did bone come from? An overview of its evolution. *Acta Orthopaedica*, 82(4):393–398, 2011.

[12] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

Chapter 2

Quantitative evaluation of evolution using comparative bioinformatics of gene co-expression networks

This chapter reviews the established methods used for analysing co-expression networks. Methods are classified into alignment-based and alignment-free methods depending on how the gene co-expression networks have their nodes compared. We explain the strengths and weaknesses of current methods and suggest potential solutions or approaches to adapt comparative analysis techniques utilized for other types of networks.

The manuscript in this chapter sets the foundations of the remaining chapters in this thesis, identifying the present limitations when comparing gene co-expression networks using current or traditional approaches. The paper motivates the development of better tools for comparing transcriptomic data, particularly for evolutionary studies. It also suggests using natural language processing (NLP)-based strategies for future work in gene co-expression network comparison, which is further explored in Chapter 5. The tools and methods proposed in this thesis are compared to the methods introduced in this review chapter when appropriate. This chapter will be submitted to *Frontiers in Genetics*.

Citation

Katie Ovens, B Frank Eames, and Ian McQuillan. 2020. Quantitative evaluation of evolution using comparative bioinformatics of gene co-expression networks. Under preparation for submission to *Frontiers in Genetics*.

Author contributions

Katie Ovens wrote the paper and generated the figures. Brian Eames and Ian McQuillan co-supervised the work and assisted with revision of the manuscript.

Abstract

Similarities and differences in the associations of biological entities among species can provide us with a better understanding of evolutionary relationships. Often the evolution of new phenotypes results from changes to interactions in pre-existing biological networks and comparing networks across species can identify evidence of conservation or adaptation. Gene co-expression networks, constructed from high-throughput gene expression data, can shed light on the coordinated activity of groups of genes, their evolution, and how they lead to new phenotypes. Perhaps surprisingly, there has been little work on these types of networks to study evolution. Most research has focused on protein-protein interaction (PPI) networks. While some PPI bioinformatic methods can be used to compare co-expression networks, they often disregard highly relevant properties, including the existence of continuous and negative values for edge weights. Also, the lack of comparative datasets in non-model organisms has hindered the study of evolution using PPI networks. In contrast, the abundance of gene expression data recently makes gene co-expression networks a valuable tool for the study of evolution in non-model organisms. In this paper, we review techniques for comparing gene co-expression networks in the context of evolution, including local and global methods of graph alignment. We also discuss limitations and challenges associated with cross-species comparison using gene co-expression networks, discuss the use of PPI network analysis methods for co-expression network analysis, and provide suggestions for utilizing gene co-expression network alignments as an indispensable tool for evolutionary studies going forward.

2.1 Introduction

Biological systems can be studied as large-scale networks such as gene expression networks, protein-protein interaction (PPI) networks, and metabolic networks [73]. Comparing these networks is valuable for understanding the relationships between biological entities across different phenotypes and throughout evolution (e.g. diseased vs. healthy, good prognosis vs. bad prognosis, mouse vs. human, etc). Studying how these networks are "re-wired" can provide more insight than studying biological entities as independent units that do not interact with each other. Many methods are available for PPI network analysis and comparison. Developing a specific PPI network is a challenging task for non-model organisms, which is critical for making evolutionary inferences [72]. Developing gene expression networks, on the other hand, is a straightforward task due to publicly available gene expression profiles for model and non-model organisms.

The relationships between genes can be inferred using an organism's transcriptome, which is the messenger RNA, or mRNA, molecules expressed by an organism. The transcriptome is closely tied to an organism's phenotype, such as morphological structure [67]; therefore, transcriptomic activity can affect organismal functions. With the advance of high-throughput technologies such as RNA-seq and single-cell RNA-seq, comparative transcriptomics has become useful for tracking gene expression changes that might underlie

molecular mechanisms of evolution [28]. Gene expression networks make it possible to study coordinated gene expression patterns across various phenotypes and organisms.

Gene co-expression networks (GCNs) represent gene–gene interactions as an undirected graph, where the nodes of the graph represent genes and an edge between two nodes represents the direct or indirect interaction between those genes [82]. Although these networks do not contain information about regulation direction, they still allow for the simultaneous analysis of many genes and the potential relationships between them. GCNs can be compared across different tissues, cell types, or species to better understand the coordinated changes in gene-gene interactions [91]. Several techniques are currently utilized to make cross-species GCN comparisons, including differential co-expression network analysis methods [4, 44, 85, 97], inter- and intra-modular hub detection [91], and functional annotation transfer [69, 70, 91].

Cross-species comparisons of GCNs can be used to detect evidence of conservation and adaptation. Past studies in comparative transcriptomics simply compared expression of mapped orthologs between species or focussed on modules that are associated with particular processes [82, 99]. Differential co-expression analysis also detects differences in the co-expressed genes between two conditions, typically diseased and healthy samples [38], but can also compare two species [91].

In this paper, we focus on using GCN comparisons of species to identify evidence of adaptation and conservation (Figure 2.1). Network alignment and alignment-free methods can address the lack of knowledge regarding how each node of one network maps to one or more nodes of the other network(s), and identify areas where GCNs are conserved or different [53]. However, several challenges exist when comparing and aligning GCNs, let alone PPI, gene regulatory, metabolic, and ontology networks. Depending on the strategy chosen, the methodology can be computationally intractable, requiring heuristics. Further, the best network alignment methods were designed to align PPI networks, so the best candidate for GCN alignment specifically is unknown.

In Section 2.2 we explain how the general representation of GCNs differs from PPI networks. In Section 2.3, we discuss the trade-offs between local, global, pairwise, and multiple alignment-based methods in the context of evolutionary studies. Section 2.4 describes the available tools and methodologies to align GCNs, including common alignment-free methods, highlighting their shortcomings. In Section 2.5, we provide suggestions for current challenges in comparing GCNs. Finally, Section 2.6 concludes the paper with a summary and conclusions.

2.2 Co-expression network representation

There are several ways in which biological networks may be represented graphically, with different methods to represent relationships between nodes. PPI networks typically have edges that have no associated weight. A weighted graph can also be used, where the edge weight can also signify how confident, based on available data or experimentation, one can be that the edge is present [27]. This is typically represented as a value

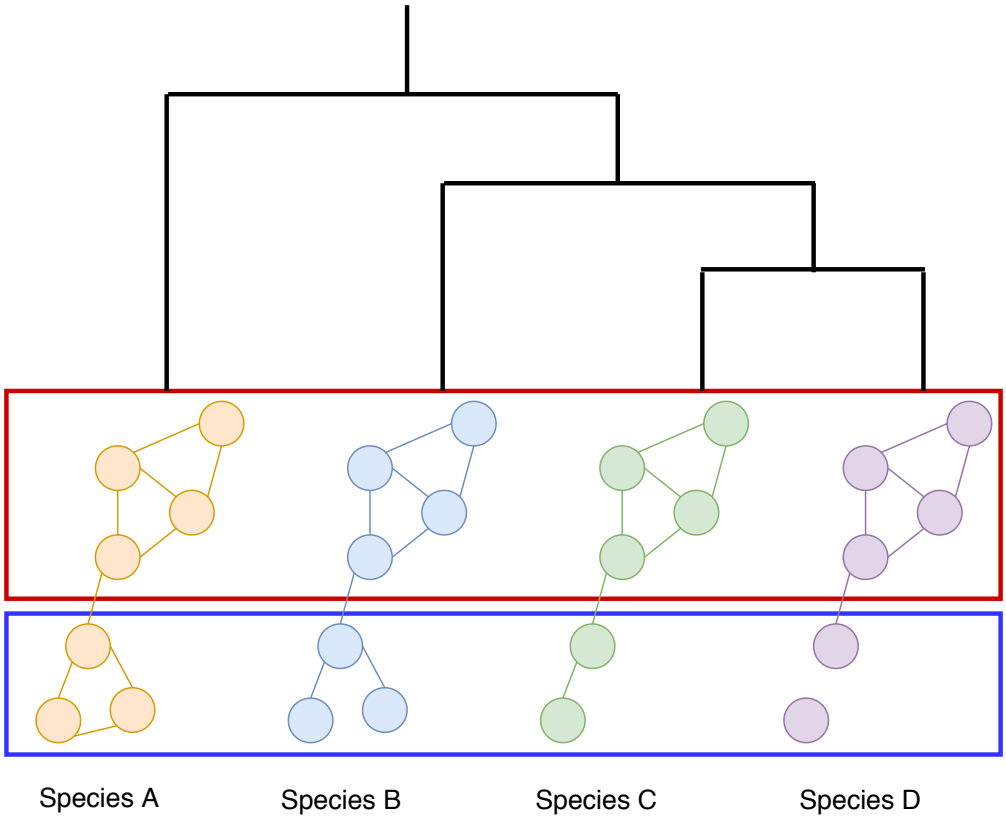
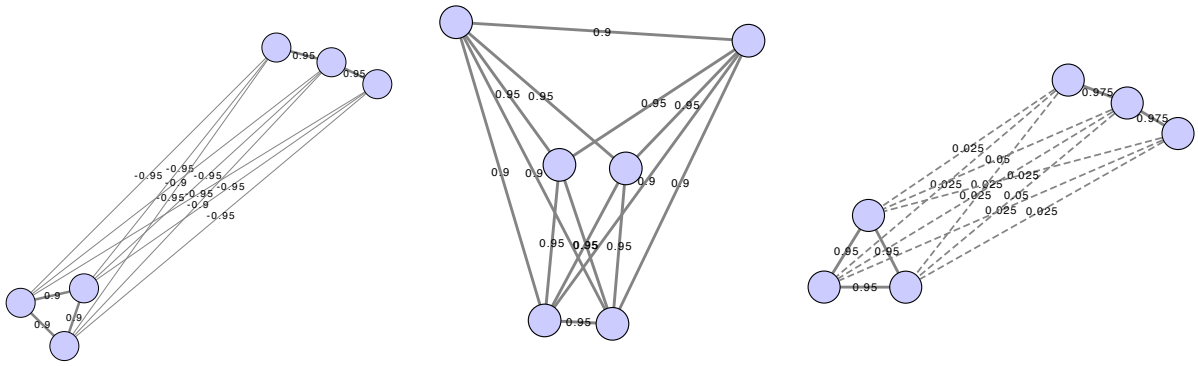


Figure 2.1: Visualization of the hypothetical changes in a biological network, such as a co-expression network, generated from 4 different species (species A, B, C, and D). The subnetworks surrounded in red represents evidence of conservation across all 4 species with all nodes and edges of each network conserved. The portion of the network surrounded in blue is evidence of adaptation, where the nodes and their connections have changed depending on when the species diverged. In this scenario, the relationships between the nodes are becoming lost over time, suggesting that perhaps whatever this module of genes was responsible for biologically is no longer important for the later diverged species.



(a) Original co-expression network (b) Unsigned co-expression network. (c) Signed co-expression network.

Figure 2.2: Co-expression networks show the difference between original (a), unsigned (b), and a signed gene co-expression networks (c). The original network shows edge weights as calculated using a correlation measure such as Pearson correlation. Networks were constructed using Cytoscape [74].

between 0 and 1, with 1 being the highest confidence and 0 being the lowest confidence. Socio-affinity index, for example, is a type of confidence score measuring the association between each pair of proteins based on an entire affinity purification-mass spectrometry dataset [25].

GCNs typically use weighted graphs. One of the most common similarity measures used to construct these weights is correlation, an association measure used to estimate the relationships between two variables. Pearson correlation coefficient measures the extent of a linear relationship between variables x and y; and Spearman correlation is based on rank, measuring the extent of a monotonic relationship between x and y. All correlation coefficients take on values between -1 and 1, where negative values indicate an inverse relationship. A correlation coefficient is an attractive association measure since it can be easily calculated, allows for calculating significance levels (p-values), and the sign (+/-) allows one to distinguish between positive and negative relationships. For gene network prediction, close relationships have been found between mutual information and correlation-based GCNs. Mutual information is often highly related to the absolute value of the correlation coefficient and when they disagree, the correlation findings appear to be more plausible statistically and biologically [79, 81]. Simple measures such as these have been found to be among the highest performing for measuring network connectivity and functional inference [5].

Some of the more common ways GCNs represent edge weights are shown in Figure 2.2. First, the edges can be weighted from -1 to 1 using simply the correlation coefficient. Alternatively, edges may be weighted using the absolute value of correlation coefficients, using

$$|cor(e_{xc}, e_{yc})|, \tag{2.1}$$

where e_{xc} is the expression of gene x in condition c. This is referred to as unsigned correlation, and has the effect of mapping both positive and negative correlation towards 1, and no correlation towards 0. Furthermore, correlation can also be transformed to be between 0 and 1 by using the following equation:

$$0.5 + 0.5 * cor(e_{xc}, e_{yc}), \tag{2.2}$$

This is referred to as signed correlation. A value closer to 0 is a strong negative correlation, a value closer to 1 is a strong positive correlation, and a value of 0.5 indicates no correlation [44]. Although this method retains information regarding negative and positive correlation, typically this method is not used to align networks. Since most of the networks aligned are designed around PPI network methods, the negative correlation would be more likely to be ignored (much as values close to 0 are ignored in PPI networks) than the high weight positive correlation using this technique. The relationship between negatively and positively correlated genes should ideally be distinguished for proper alignment. As GCNs have some additional variations in the interpretation of their edges compared to many PPI networks, low edge weights in GCNs will not have the same meaning as low edge weights in PPI networks.

These networks are often thresholded either using a strict cut-off that is applied to filter out non-important edges, filter edges based on rank—where a percentage of the most highly co-expressed genes is retained for each gene in the network—or the network has a soft threshold applied [44, 59, 89, 99]. A soft threshold retains all of the edges, but takes the edge weights to a power that makes the GCN scale-free. As a result, the lower weights are pushed closer to zero and stronger correlations are emphasized.

2.3 Graph alignment in biological networks

The principle behind biological network alignment is that biologically relevant associations are likely to be independently observed in different individuals, species, tissues, or conditions whereas false associations are less likely to be repeatedly observed. For example, the conserved genes in terms of both sequence and expression among multiple species are expected to play a key role in biological responses [82]. The goal is therefore to align the networks to identify these conserved elements. In order to better understand the application of network alignment to co-expression networks, it is important to consider the techniques used with other types of biological networks. Early work on biological network alignment has been focussed primarily on protein-protein interaction (PPI) networks [65].

For network alignment, the basic problem is represented as follows: each network is represented as a graph G_i, where $G_i = (V_i, E_i)$ with V_i being a set of nodes and E_i being the set of edges that connect nodes in V_i. Some scoring scheme is defined between components of the graph, and the goal of an alignment between two networks G_1 and G_2 is to map as many nodes and edges in one graph to the nodes and edges (respectively) of the other in such a way that the sum of scores is high. However, there are many factors that can be integrated into the scoring scheme, which will be explored next.

Network alignment strategies can be considered global or local. The goal of local alignment is to find conserved subnetworks in a graph; since multiple local alignments can exist, this means that individual nodes in one graph can have multiple good local alignments. These methods tend to identify subnetworks or communities of related genes. In comparison, global alignment methods typically align every node in one network to a node in another network, attempting to find the one alignment with the maximum amount of

similarity [55].

Network alignments can also be independently divided into two categories: uniquely labelled, and unlabelled. For the first, the two graphs have labelled nodes, which could be e.g. gene name (in principle, graphs could have labelled edges as well). A uniquely labelled network has separate labels for each node. In a uniquely labelled alignment, it forces a node to align with the (at most one) similarly-labelled node in the other graph. It should be noted that it is possible to create optimal uniquely labeled alignments in a computationally efficient manner (in polynomial time) [16]. An example of an alignment of uniquely labeled networks maps only one-to-one orthologs between species to each other. In contrast, unlabelled alignments ignore any labels on the graphs, and match based on topological similarity only. Both unlabelled local and global graph alignment are usually computationally intractable to solve optimally. As an example, just the problem of determining whether two graphs are isomorphic (they are the same after renaming nodes and edges) has no known polynomial time algorithm. Furthermore, the subgraph isomorphism problem takes two unlabelled graphs as input and attempts to determine whether the first has a subgraph that is isomorphic to the second graph. This problem, and its generalizations, are known to be NP-complete and therefore likely intractable [24]. Therefore, heuristics need to be applied in order to find the areas that match (or are topologically similar) between networks.

It is also possible to utilize a cost function that combines the topological and/or homology similarity between the nodes of each network. For example, one such cost function is

$$C(u_i, v_j) = \alpha T(u_i, v_j) + (1 - \alpha) H(u_i, v_j), \tag{2.3}$$

where v and u are nodes in network i and j, and T is some topological score, representing the similarity of the topological neighbourhood of the nodes in their networks; and H is a homology score indicating the similarity of the genes at a sequence level. This means that the alignment of orthologs could be considered in the calculation of a score, but it does not necessarily enforce a mapping between these orthologs. This can be thought of as a hybrid of the uniquely labelled and unlabelled approaches. In order to vary how much influence each of the similarities have to the overall score, a parameter α is used in Equation 2.3, which is a fixed value between 0 and 1. The closer α is to 0, the more influence homology has on determining the similarity between nodes; the closer α is to 1, the more influence comes from topological information. Some of the newer methods of network alignment also allow for updating this cost function after each iteration of an alignment—after some nodes have already been aligned—which could provide information for the remaining iterations [30, 62, 83].

The following sections will briefly describe work in comparing networks using homology and topology as similarity measures. Most of the specific examples of alignments involve PPI networks as this is the biological data for which most of these methods have been designed. Comprehensive reviews of the many methods or tools available for network alignment are covered in the literature [13, 18–20, 30, 55].

2.3.1 Homology similarity in biological network alignment

An important aspect of generating informative network alignments is ensuring they are biologically accurate. Homology is usually identified using BLAST between protein sequences with E-values less than some threshold, or by identifying and working with the known orthologs between species [82]. Some methods use these calculated BLAST E-values as part of their alignment cost functions, or can incorporate a variety of similarity information including sequence, structural, and ontology information [13].

Utilizing Gene Ontology (GO) is another strategy for not only evaluating, but also sometimes driving the alignment of networks [43]. GO terms are a controlled vocabulary that describes biological properties of gene products, and the Gene Ontology is the organization of these terms and describes their relation to each other. The terms are organized as a directed acyclic graph (DAG) where each node is a GO term, and each edge is the relationship between the GO terms. One strategy is to determine the semantic similarity between the GO terms for each node across the networks being compared [76]. To do so, the subgraph of GO terms annotating each node of a network is transformed into a vector of information content distance for every GO term pair. A pair of nodes across the networks being compared can then be compared, and a measure of similarity between the nodes is calculated as the Euclidean norm between the distance vector for each node to get a similarity score and determine good alignments between the networks. Another simple method typically used to evaluate an alignment based on GO terms is to calculate the fraction of aligned proteins sharing the same GO terms [43]. The larger the fraction, the more biologically meaningful the alignment. The GO terms can also be weighted based on their frequency or how informative they are [32].

Using homology alone as a means to align or compare networks is sometimes limited by the amount of detailed annotation available for each species. Furthermore, using annotation to align the networks likely will not be as useful if the goal is to align networks in order to transfer annotation from one species to another. Therefore, incorporating topological information is also useful for guiding network alignments.

2.3.2 Topological similarity in biological network alignment

Some alignment methods rely on strategies to measure similarity between the topological properties of networks. Common similarities include calculating differences between degrees, clustering coefficients and eccentricities [31, 42], spectral signatures [47, 68, 78], and graphlet-degree signatures [50, 53, 58]. For example, alignment could involve aligning graphs based on similarity of neighbours, where two nodes are considered a good match if their neighbours are also good matches.

IsoRank is the original graph alignment method introduced to align PPI networks [78], and it has also been used to align GCNs [21, 47, 99]. In the original algorithm, the guiding principle was that if two nodes of different networks are aligned, then their neighbours should be aligned as well. It is an application that uses spectral methods, whereby the eigenvalues and the eigenvectors of the adjacency matrix of a graph are invariant with respect to node permutations. Therefore, if two graphs are isomorphic, their adjacency

matrices will have the same eigenvalues and eigenvectors [14]. IsoRank uses this spectral graph theory result to build a multiple network alignment by local partitioning of the graph of pairwise functional similarity scores, which are calculated for every pair of cross-species proteins. Next, a greedy algorithm is used to produce the alignment. An updated method, IsoRankN, does not use the greedy algorithm, and instead uses an iterative spectral clustering algorithm. A similarity graph is constructed with a protein set for each species and edges connecting them are weighted by a similarity score [47]. The highly weighted edges and the neighbours connecting them are identified, and the total weight is calculated. The proteins are then ordered by their total weights using an iterative spectral clustering algorithm to identify the conserved proteins. IsoRank and IsoRankN are capable of aligning 5 and 6 species at most, respectively, due to their exponential time complexity [33]. Furthermore, handling large networks of more that 10,000 proteins or genes is a challenge [75].

2.3.3 Current directions to improve biological alignment strategies

The main strategies for improving alignment methods are to (1) combine local and global alignment methods [54, 57], (2) improve the agreement between topology and homology similarity [30], (3) consider both node and edge similarities when making alignments [15, 83, 93], (4) align more than two networks [22, 41, 47, 92], and (5) combine groups of alignment methods [49, 51]. The limitations of using either local or global alignment is being addressed with methods that try to find a balance between local and global alignment, which have been shown to be complementary [54]. Therefore, it may be beneficial to use both for any study or incorporate features of both alignment methods in a single method. IGLOO, for example, utilizes an already available (interchangeable) local alignment method to make an initial alignment, and then applies a global alignment strategy to improve topological similarity [54]. As another example, GLAlign initially applies MAGNA++ (a global alignment method) to collect a list of matching nodes and a list of seed nodes generated from biological information. Then Align-MCL (a local alignment method) is used to produce the final alignment [57].

The majority of the methods described in this section have only been tested with and applied to PPI networks. For making evolutionary inferences from other biological networks, the ability to align many species using a multiple network alignment method would also be useful. The following section describes how GCNs have been compared using network alignment including methods that have been applied or designed specifically with GCNs in mind.

2.4 Alignment and alignment-free methods and applications to gene co-expression networks

The benefit of using gene expression as opposed to PPI networks is that the PPI networks available today across a variety of species are largely incomplete. Depending on the species or tissues a researcher wishes

to study, it may be difficult to obtain enough PPI information. It is much easier to collect high-throughput sequencing data for many species, which can be used to generate GCNs. From an evolutionary perspective, these networks can be used to identify likely functional orthologs in species with less information, identify evolutionarily conserved sub-graphs, as well as identify conserved functions.

Co-expression networks exhibit many of the same properties as PPI networks. They both tend to have a scale-free structure and have a strong modularity [9]. They also both have a number of highly connected nodes that are known as hubs [23]. However, although many GCNs have been constructed, few PPI network alignment techniques have been utilized for comparing GCNs, especially from eukaryotic organisms. Section 2.4.3 contains a discussion of PPI alignment methods that have been applied to GCNs, methods developed specifically to align GCNs, and a description of their applications. First, Section 2.4.1 describes some methods and applications of comparing GCNs without creating alignments.

2.4.1 Alignment-free comparisons of co-expression networks

Alignment-free network comparisons aims to quantify the similarity between networks [60, 61, 73, 90]. Among these approaches are measuring the similarity between the topological properties of networks [2, 46, 59, 89], clustering for the identification of conserved modules of genes [26, 82, 99], and comparison of edge weights for matched orthologs [39]. Since these methods are not designed to (directly) generate a mapping between the nodes of the networks, beyond the known orthologous relationships, we do not consider them as network alignment methods. However, many of these methods work to match up groups of genes, or clusters, so we discuss these methods in Section 2.4.1. Section 2.4.2 includes a demonstration of WGCNA, which is a commonly used method for comparing GCNs.

Cluster alignment methods

Clustering has been utilized to identify evidence of conservation in gene co-expression across vertebrate species [10, 26, 66, 98]. Many methods designed explicitly for co-expression network comparison generate a mapping between clusters [99]. These methods link modules of co-expressed genes together based on the known orthology relationships of genes. We refer to these methods as cluster alignment methods.

Yan et al. proposed OrthoClust based on a simulated annealing strategy. OrthoClust aims to discover the optimal assignment of orthologs to modules based on a cost function considering the modularity and known orthologous links between genes within clusters [99]. They evaluated their method based on a set of 1288 genes reported to have conserved expression patterns across several species, including worm and fly. These genes were referred to as metagenes and expected to be in aligned clusters. The authors reported that when compared to the alignment method IsoRank, 88% of metagenes were aligned by IsoRank while 81% were grouped in the same clusters by OrthoClust. This observation suggests that PPI network alignment methods could lead to biologically meaningful results for comparing GCNs.

A limitation of most clustering-based approaches is that they assign each gene to a single cluster; however,

genes could be involved in different regulatory pathways depending on the conditions they are acting under. Biclustering on the other hand, can be used to simultaneously cluster genes and samples to detect co-expressed genes under different subsets of conditions [96]. Each module of genes or bicluster could have co-expressed genes under different subsets of conditions, and genes may be contained in multiple modules. The biclusters identified can be used to predict subnetworks more quickly than trying to construct a network using all of the genes at once, which can later be merged into a single network by removing duplicate edges [1]. Biclustering can also be utilized to compare gene expression across tissues and species. For example, it has been applied to 50 different human tissues to identify gene relationships particular to individual tissues [71]. The application of biclustering to identify conserved and unique gene expression patterns across different species has been limited [34, 40, 94].

COMODO uses adaptive co-clustering to compare up to 3 species [100, 101]. The algorithm starts with a gene–gene correlation matrix where each axis of the matrix is for one of two species, and genes that are co-expressed more highly are grouped together in modules at a specified threshold, which is determined using biclustering [7]. The groups below the diagonal entries in the matrix that are locally more co-expressed with each other than with their neighbouring genes are considered the seed modules. These seeds are expanded in each species until a pair of modules is obtained for which the number of shared orthologs is statistically optimal relative to the size of the modules. Module seeds linked by a sufficient number of orthologous gene pairs are gradually extended by traversing the space of possible cluster threshold combinations, using a combination of greedy and brute force search, represented on the gene–gene threshold matrices of each species until optimality is reached. These comparison techniques appear to have several drawbacks. First, the method of evaluation relies on the quality of functional annotation available for each species. Also, multiple cut-offs may need to be applied to determine the best co-expression stringency values for identifying possible seed modules. As modifications were also made to reach stopping criteria more quickly and reduce the memory required, it is questionable whether this method could be further extended to compare more than 3 species. Lastly the researchers explain that the species they compare have genes that have one or two corresponding homologs in the other species, which is required for their method to work as expected [101]. Therefore, if the species compared are evolutionarily distant, or have a large portion of one-to-many or many-to-many mappings, using their statistic may not be possible.

Clustering and biclustering are useful strategies to reduce the dimensions of gene expression data. Both of these strategies can be used to identify modules of genes, which can be utilized for functional analyses or comparisons between the identified modules [71]. Below we discuss and demonstrate one of the more common strategies used to construct and compare GCNs that utilizes clustering as one part of their analyses.

2.4.2 WGCNA for comparing gene co-expression networks

One of the most widely used techniques to compare gene expression datasets is to use weighted gene co-expression network analysis (WGCNA). Although WGCNA was created in 2008, it is still commonly used to

detect potentially important modules of genes associated with diseases [3, 84], biological pathways [77], and development [80]. First, unsigned or signed correlation is calculated using Equation 2.1 or 2.2, respectively. These values are used to construct the adjacency matrix, which is a quantitative measure of the strength of the relationship between each pair of genes. Each value of the adjacency matrix is raised to a power β, which is the smallest value of β that can be used where a scale-free topology is achieved. Next, WGCNA uses a topological overlap measure (TOM), which is a combination of the adjacency value between a pair of genes as well as the adjacency values these genes have with other genes they are connected to.

Methods like WGCNA and other differential co-expression analysis methods can be used to identify conserved clusters as well as clusters that contain different genes or behave differently under changing conditions or phenotypes. WGCNA offers module preservation statistics to make comparisons across modules of different clusterings. In order to measure the preservation of a module, WGCNA can be used to determine if it is reproducible (or preserved) in an independent test network. One score is Zsummary score, which is a composite score of density and connectivity preservation statistics to determine if a module is significantly more similar to a reference module than a random sample of genes [45]. As module size dependence could be an issue, medianRank can also be calculated for each module, which is a rank-based measure of the density and connectivity statistics. Each module is ranked based on the observed values for the statistics for each module.

Figure 2.3 shows the results of WGCNA applied to publicly available RNA-seq datasets in human and macaque from Bozek et al. [8]. Samples from the prefrontal cortex were used to construct the GCNs. Relatedness of network connectivity to molecular rates of evolutions has also been studied by using WGCNA and estimates of dN (nonsynonymous substitutions per site) and dS (synonymous substitutions per site) [48, 52]. More highly connected genes or genes found in a greater number of cross-tissue modules showed greater sequence constraint.

As the majority of differential co-expression analysis methods only focus on comparing two biological states, BioNetStat [37] can be used to compare various topological features such as degree, eigenvector, closeness, betweenness, and clustering coefficient distributions between any number of networks to discover alterations. However, although this method can tell if there is a topological difference between networks, how the genes participating in these interactions have changed is less clear; i.e., it is possible to tell if degree centrality for a specific gene has changed, but not how many genes it is connected to have changed. This method has yet to be tested on multiple species, but since it is capable of making pairwise measurements between multiple networks, it may be a candidate method for studying evolution.

Network alignment alternatively tries to find the node correspondence between networks that leads to highly similar conserved network regions. Both approaches have their own set of challenges and depending on the biological question of interest, and how well characterized a species is, one approach may have advantages over the other.

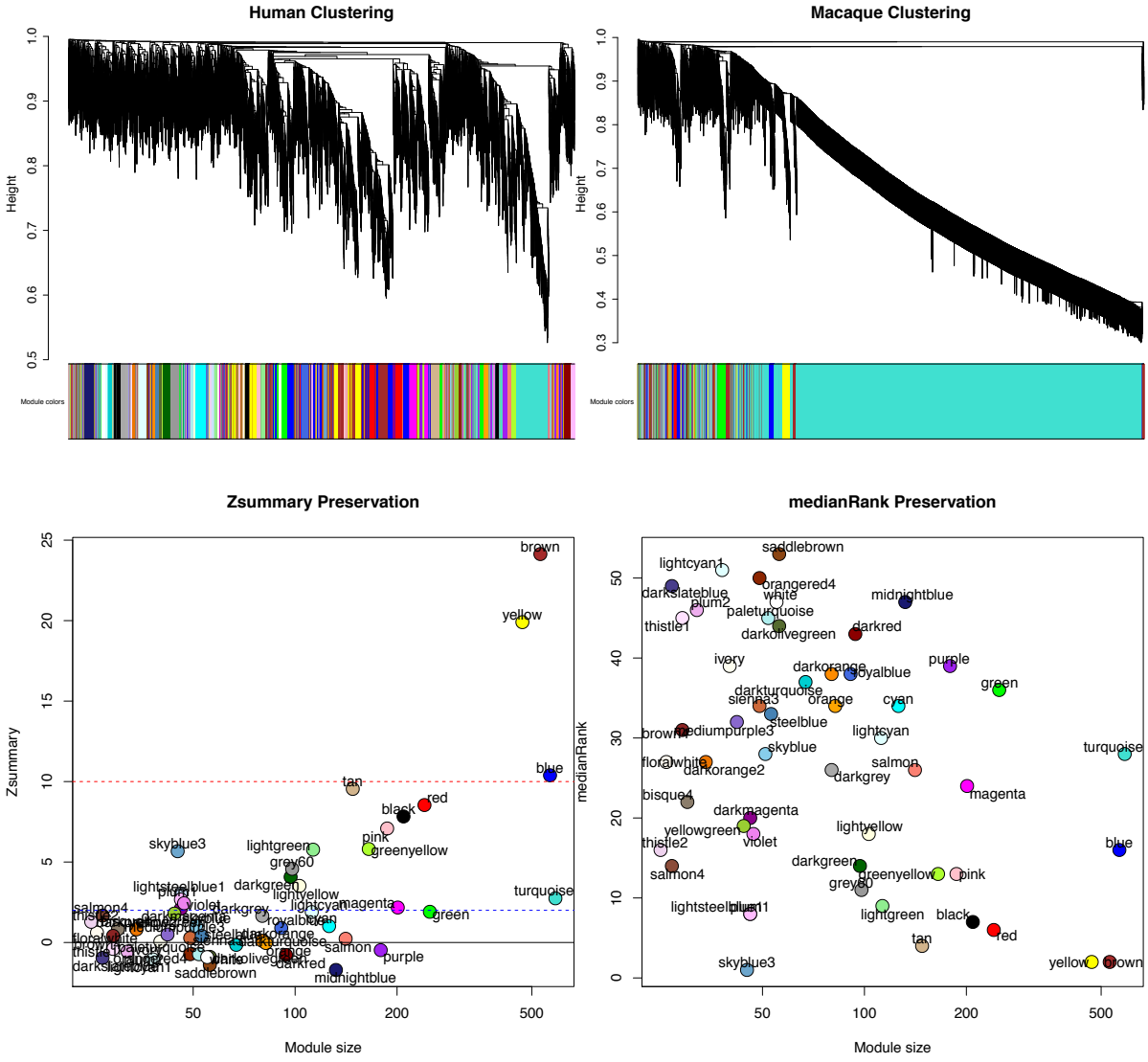


Figure 2.3: Module preservation statistics comparing signed gene co-expression networks constructed using prefrontal cortex samples from human and macaque. To generate the networks, a power of $\beta = 8$ for soft thresholding was applied to create a scale-free network topology. The clustering merge height was set to 0.20 to generate the clusters shown in the dendrogram and module colour images (bottom). The minimum module size allowed was 30 genes. From both Zsummary (top left) and medianRank (top right) preservation scores using mouse as the reference network, the most preservation is observed for the yellow and brown modules. A Zsummary score greater than 2, but less than 10 indicates moderate preservation, while a score greater than 10 indicates strong module preservation. A low score for medianRank indicates high module preservation. One limitation of Zsummary score is that it often shows a dependence on module size meaning larger modules tend to get a higher score. It is also computationally intensive as it relies on permutation tests to determine significance. Although medianRank is not module size dependent like Zsummary, one drawback of medianRank is that it is rank based and therefore, it can only measure relative preservation. For example, the yellow and brown module with low medianRank scores may not be that well preserved, but it is the most preserved in comparison to the other modules discovered.

Table 2.1: Studies utilizing an alignment strategy to compare GCNs

Organism	Author	Year	Method/Technique	Sample description	Alignment type
Rice	Ficklin and Feltus	2011	IsoRankN	508 microarray samples	global, pairwise
Maize	Ficklin and Feltus	2011	IsoRankN	253 microarray samples	global, pairwise
Arabidopsis	Wang et al.	2013	Subnetwork alignment	Leaf, flower, shoot microarray samples	local, pairwise
Poplar	Wang et al.	2013	Subnetwork alignment	Leaf, flower, shoot microarray samples	local, pairwise
Mouse	Berg and Lässig	2006	Bayesian alignment	61 tissues	global, pairwise
Human	Berg and Lässig	2006	Bayesian alignment	79 tissues	global, pairwise
Mouse	Wang et al.	2009	SCHype	300 microarray liver samples	global/local, pairwise
Human	Wang et al.	2009	SCHype	423 microarray liver samples	global/local, pairwise
Rat	Wang et al.	2009	SCHype	382 microarray liver samples	global/local, pairwise
Mouse	Towfic et al.	2010	BiNA	45 tissues, organs, and cell lines, 90 microarray samples	local, pairwise
Pig	Towfic et al.	2010	BiNA	16 tissues, 64 microarray samples	local, pairwise
Human	Towfic et al.	2010	BiNA	46 tissues, organs, and cell lines, 85 microarray samples	local, pairwise
Mouse	Towfic et al.	2012	BiNA	33 ligands, 422 microarray B-cell samples	local, pairwise
Fly	Yan et al.	2014	IsoRank	30 developmental stages RNA-seq samples	global, pairwise
Worm	Yan et al.	2014	IsoRank	33 developmental stages RNA-seq samples	global, pairwise
Fly	Nguyen et al.	2019	ManiNetCluster	12 timepoints RNA-seq samples	global, pairwise
Worm	Nguyen et al.	2019	ManiNetCluster	25 development stages RNA-seq samples	global, pairwise

2.4.3 Alignment-based methods and applications to gene co-expression networks

Table 2.1 shows GCN alignments that have been published in literature. Few graph alignment methods have been described from a GCN perspective or utilized to compare GCNs across different species to make inferences about their evolution.

Ficklin and Feltus [21] utilized IsoRankN [47], designed for PPI network alignment, to compare GCNs constructed from rice and maize. The focus was to transfer functional annotation from maize to the less-characterized rice GCN. They identified 194 genes that had unknown function in rice through 3,092 conserved edges, which suggested associated biological processes such as seed storage. Interestingly, although sequence orthology in general is a common strategy for transferring functional annotation from one species to another, the cost function used to generate the alignment between these species was weighted with more emphasis towards topological similarity. This suggests that similarities in topological structure between GCNs is informative for functional annotation transfer.

A study of *Arabidopsis* and *Poplar* incorporated analysis of network topology and also went so far as to align the networks to identify the conserved and species-specific functions of cell-wall related genes [95]. Subnetworks associated with cell wall genes in leaf, flower, and shoot tissues between the two plant species were aligned while considering the neighbouring orthologous genes. Tissues that had good alignments were considered to likely have more conserved function. They also separately investigated network centralities including clustering coefficient and eigenvector centrality for measuring a gene's global influence over the entire network. Conserved hub genes and tissue-specific hub genes across networks were discovered.

Berg and Lässig [6] utilized a probabilistic alignment procedure for biological network alignment based on their edge and node similarity and attempted to maximize their proposed score based on a mapping to a generalized quadratic assignment problem. This was another method proposed to identify conserved modules of genes, which in this case was applied to compare human and mouse GCNs. However, they applied their method to a limited number of genes considered housekeeping genes that were expressed in all samples and showed a low variance of expression levels across samples in both species, as well as genes with a high expression similarity with at least one of the genes considered housekeeping genes. Furthermore, although they claim to analyze the evolution of GCNs between humans and mice, studying evolution is challenging given that only two species were compared.

SCHype addresses local and global alignment with recursive spectral clustering and biclustering algorithms of hypergraphs (generalizations of graphs where the edges can exist between arbitrary subsets of nodes, rather than just two) to identify sets of nodes in each species with a greater than expected number of conserved interactions (based on co-expression in this case) between them [56]. The technique is used to discover densely interconnected genes by computing the dominant eigenvector and then converting it to a discrete set of vertices. Uniqueness of the dominant eigenvector guarantees unambiguity of the solution and rapid convergence of the procedure and allows for analysis of complex homology groups, unlike the potential drawbacks of using Pearson's chi squared test from COMODO [100]. Again, this is another technique that has been applied to GCNs for the purpose of functional annotation transfer [70]. Importantly, individual experiments from the series did not cluster together in SCHype clusters. This could indicate different normalization techniques may be required to account for these differences to make better comparisons across experiments.

BiNA is another PPI network alignment method that has been applied to co-expression data, and works by breaking down the networks being compared into subgraphs to align each of them [86, 87]. A "k-hop subgraph" for each gene g is constructed by including any gene that can be reached from g within k edges. Each of the subgraphs in one co-expression network is compared to the subnetworks in the second network to find the best match based on a similarity measure, taking into account the sequence homology of the genes as well as the general topology of the neighbourhood around each gene. The authors used their method for ortholog detection [87], and to identify B-cell ligand processing pathways [86]. None of the similarity measures employed by this method have utilized weighted edges, but there is potential to extend the method to handle weighted graph scenarios.

ManiNetCluster is a recent strategy for alignment that projects two GCNs into a common lower dimensional space on which the Euclidean distances between genes preserve the geodesic distances between them [63]. This distance was used as a metric to detect manifolds embedded in the original high-dimensional network space. This is different from the other alignment methods covered in that it is a subspace learning approach, embedding the nodes across different networks into a common low dimensional representation. The network representations were used to form a multilayer network that could be clustered to a number of

cross-network gene modules. The benefit of such a method is that only the networks are required to learn the underlying structure of the data. This method was only tested on a cross-species comparison using 1882 fly genes and 1925 worm genes. Therefore, it would need to be tested on a much larger network with more species to determine its utility for making evolutionary inferences.

These methods have not been systematically compared with other network comparison strategies so it is not clear to what effect aligning the networks has on detecting evidence of conservation or adaptation. From Table 2.1, it is also clear that RNA-seq has not been highly utilized to perform network alignments although there are many instances of RNA-seq being used to construct, analyse, and evaluate GCNs in other ways [5, 35]. Adapting network alignment methods to work with GCNs comes with several challenges and some of these challenges may explain why there is little overlap between the methods used to compare PPI networks, and the methods to compare GCNs. In the following section, the possible limitations that may prevent PPI analysis methods from being commonly used for GCN analysis are described.

2.5 Challenges and future directions

Using gene co-expression data for network analysis and alignment has some advantages over PPI network analysis and alignment, such as the much larger availability of data for the study of transcriptomics, but it also has some limitations. Gene co-expression cannot provide a full understanding of complex gene-gene interactions because they cannot distinguish between direct and indirect interactions. In other words, if they are viewed as networks that only contain direct, causative, and directional interactions, GCNs can contain many false positive interactions and the interpretation of evolutionary rewiring by focussing on specific interactions is more limited. This is why co-expression network analysis tends to focus on changes that are occurring in groups or modules of genes. GCN network alignment is an under-utilized tool for identifying conserved subnetworks across multiple species to study evolution. Some of these methods do not require knowledge beyond the GCNs such as orthologous genes, which is also useful for studying non-model organisms where there can be more unknown functional links between genes.

The sign of the edge weights connecting nodes of a network can mean different things depending on whether the network is a PPI or GCN. Depending on how often differences in the edge weight sign are observed when comparing the relationship between genes in two species, it may be important, or negligible. If the relationship between genes changes between species, it is possible that the relationship identified is not a direct or conserved relationship. Some information on these relationships would be lost using different methods of calculating co-expression if they are modified in a way that treats negative and positive correlations equally. Also, aligning GCNs using PPI strategies may confound these relationships between genes. However, it may be possible to align the networks by incorporating the change in correlation between each pair of genes across a pair of species into the score to increase biological accuracy.

Gene co-expression networks tend to be much larger than PPI networks, with many more edges. Alignment-

based methods for comparing networks can fail to find a good alignment between these networks. Even when trying to align a large PPI network to itself, many alignment methods can report low edge and/or node conservation [36]. Using large numbers of samples may reduce the number of false positive edges in GCNs, but depending on the thresholds used to decide what edges should be included, there is still a large number of edges to consider.

Another possible limitation is the sample size of the dataset used to construct a network [5] and finding multi-species studies to make evolutionary inferrences. As it is often impractical to expect large datasets to be generated containing many species, it would be beneficial to make use of other publicly available datasets. However, this can result in technical challenges where network structure is determined in part by data biases. Although batch normalization methods are available, there are few normalization methods to address differences between environmental conditions [64]. For example, not all species may be sequenced by the same lab or have different conditions in which they are raised and bred. As such, these different conditions are likely to have an impact and need to be accounted for to prevent misinterpretations. Therefore, a comparative method to uniformly analyze cross-condition or cross-species gene expression data is essential.

The majority of studies identified in Section 2.4 only consider 2 or 3 species when utilizing alignment-based methods to compare GCNs. Therefore, it is challenging to make any inferences about the evolution of genes and the processes they drive as more than 2 species are required in order to provide evolutionary trajectory. Many of the current methods of comparing GCNs are not designed to handle more than 2 comparisons. One reason these studies limit the number of species could be that the methods that rely in part on homology to make comparisons may only be appropriate if the sequence divergence between the species compared is sufficiently small so that all pairs of functionally related nodes can be mapped by sequence homology. However, genes with entirely unrelated sequence may take on a similar function in different organisms, and hence have a similar position in two networks. As such, currently little is inferred from an evolutionary perspective when comparing GCNs.

Furthermore, identifying evidence of adaptation across GCNs is rarely the focus of alignments. As heuristics are used with the goal of identifying areas of conservation in the networks, it may not imply that what is not identified as conserved should be considered evidence of adaptation. As an example, in the alignment of maize and rice, anything not considered conserved was discarded [21]. It is possible that incorporating information on how much variation these genes have within each species will help to determine if areas of the networks not considered conserved were left out because of the greedy nature of an alignment algorithm, or if it is actually strong evidence of adaptation. It may also be possible to utilize differential co-expression algorithms that are capable of comparing more than 2 species to identify areas of the networks that are most likely changing. Despite these limitations, GCNs have the potential to provide valuable glimpses into complex gene-product interactions, especially if the information can be combined with other biological networks.

Since graph alignment, in general, has been utilized for so long [13, 18–20, 30, 55], application of more of these methods to GCNs may be a good first step before attempting to create new alignment methods

specifically for GCNs. PPI networks, for example, have been utilizing methods to align tagged social networks [56, 102]. At the very least, GCN aligners should be systematically compared to other PPI alignment methods to show how they are suited for this task. As IsoRank has not performed very well in PPI network alignment based on evaluation studies [49], it may be beneficial to adopt others that have performed better to make alignments in the future.

Finally, as GCN structure tends to be difficult to compare, one possibility for future research in cross-species GCN analysis is to utilize embedding strategies, typically used in natural language processing to generate numerical representations for genes. Traditional techniques such as matrix factorization have shown promising results, as well as more recent random walk-based and neural network-based methods [29]. Embeddings are frequently faster than other options that operate on the original networks and are less sensitive to structural noise compared to spectral methods [88]. Additionally, the learned embeddings are often applicable for downstream analysis by direct interpretation of the embedding space. Co-expression networks have recently been used to generate gene representations for single networks [11, 12, 17] and a manifold learning technique has been used to compare co-expression networks [63]. This may be an avenue of research for comparing an increasing number of biological networks in the future with improved and state-of-the-art techniques now available for embedding in natural language processing research.

2.6 Conclusion

Methods to compare gene expression among species include GCN alignment, which can identify quantitative evidence of adaptation or constraint acting on various groups of genes among species. The techniques used to align biological networks are continually being improved upon to increase the agreement between topological and homology measures of network similarity.

Graph alignment techniques have been available for a long time and used for many different applications, so we reviewed how network alignment has been applied to GCNs, highlighting any crossover with PPI alignment techniques. As the alignment of GCNs becomes increasingly common, other research areas outside of biological research might provide insights. New network comparison techniques should be enlisted to compare GCNs in more organisms with the increase in transcriptomic data from newer high-throughput technologies.

References

[1] Fadhl M Alakwaa, Nahed H Solouma, and Yasser M Kadah. Construction of gene regulatory networks using biclustering and Bayesian networks. *Theoretical Biology and Medical Modelling*, 8(1):39, 2011.

[2] Waqar Ali, Tiago Rito, Gesine Reinert, Fengzhu Sun, and Charlotte M Deane. Alignment-free protein interaction network comparison. *Bioinformatics*, 30(17):i430–i437, 2014.

[3] Mariet Allen, Xue Wang, Jeremy D Burgess, Jens Watzlawik, Daniel J Serie, Curtis S Younkin, Thuy Nguyen, Kimberly G Malphrus, Sarah Lincoln, Minerva M Carrasquillo, et al. Conserved brain myelination networks are altered in alzheimer's and other neurodegenerative diseases. *Alzheimer's & Dementia*, 14(3):352–366, 2018.

[4] David Amar, Hershel Safer, and Ron Shamir. Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Computational Biology*, 9(3), 2013.

[5] Sara Ballouz, Wim Verleyen, and Jesse Gillis. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*, 31(13):2123–2130, 2015.

[6] Johannes Berg and Michael Lässig. Cross-species analysis of biological networks by Bayesian alignment. *Proceedings of the National Academy of Sciences*, 103(29):10967–10972, 2006.

[7] Sven Bergmann, Jan Ihmels, and Naama Barkai. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review E*, 67(3):031902, 2003.

[8] Katarzyna Bozek, Yuning Wei, Zheng Yan, Xiling Liu, Jieyi Xiong, Masahiro Sugimoto, Masaru Tomita, Svante Pääbo, Raik Pieszek, Chet C Sherwood, et al. Exceptional evolutionary divergence of human muscle and brain metabolomes parallels human cognitive and physical uniqueness. *PLoS Biology*, 12(5):e1001871, 2014.

[9] Marc RJ Carlson, Bin Zhang, Zixing Fang, Paul S Mischel, Steve Horvath, and Stanley F Nelson. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, 7(1):40, 2006.

[10] Esther T Chan, Gerald T Quon, Gordon Chua, Tomas Babak, Miles Trochesset, Ralph A Zirngibl, Jane Aubin, Michael JH Ratcliffe, Andrew Wilde, Michael Brudno, et al. Conservation of core gene expression in vertebrate tissues. *Journal of Biology*, 8(3):33, 2009.

[11] Jonghwan Choi, Ilhwan Oh, Sangmin Seo, and Jaegyoon Ahn. G2Vec: Distributed gene representations for identification of cancer prognostic genes. *Scientific Reports*, 8(1):13729, 2018.

[12] Chi Tung Choy, Chi Hang Wong, and Stephen Lam Chan. Embedding of genes using cancer gene expression data: biological relevance and potential application on biomarker discovery. *Frontiers in Genetics*, 9:682, 2018.

[13] Connor Clark and Jugal Kalita. A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics*, 30(16):2351–2359, 2014.

[14] Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(03):265–298, 2004.

[15] Joseph Crawford and Tijana Milenković. Great: graphlet edge-based network alignment. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 220–227. IEEE, 2015.

[16] Peter J Dickinson, Horst Bunke, Arek Dadej, and Miro Kraetzl. Matching graphs with unique node labels. *Pattern Analysis and Applications*, 7(3):243–254, 2004.

[17] Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, 20(1):82, 2019.

[18] Ahed Elmsallati, Connor Clark, and Jugal Kalita. Global alignment of protein-protein interaction networks: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(4):689–705, 2016.

[19] Frank Emmert-Streib, Matthias Dehmer, and Yongtang Shi. Fifty years of graph matching, network alignment and network comparison. *Information Sciences*, 346:180–197, 2016.

[20] Fazle E Faisal, Lei Meng, Joseph Crawford, and Tijana Milenković. The post-genomic era of biological network alignment. *EURASIP Journal on Bioinformatics and Systems Biology*, 2015(1):3, 2015.

[21] Stephen P Ficklin and F Alex Feltus. Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. *Plant Physiology*, 156(3):1244–1256, 2011.

[22] Jason Flannick, Antal Novak, Chuong B Do, Balaji S Srinivasan, and Serafim Batzoglou. Automatic parameter learning for multiple local network alignment. *Journal of Computational Biology*, 16(8):1001–1022, 2009.

[23] Chris Gaiteri, Ying Ding, Beverly French, George C Tseng, and Etienne Sibille. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. *Genes, Brain and Behavior*, 13(1):13–24, 2014.

[24] Michael R Garey and David S Johnson. *Computers and Intractability*, volume 174. Freeman San Francisco, 1979.

[25] Anne-Claude Gavin, Patrick Aloy, Paola Grandi, Roland Krause, Markus Boesche, Martina Marzioch, Christina Rau, Lars Juhl Jensen, Sonja Bastuck, Birgit Dümpelfeld, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006.

[26] Mark B Gerstein, Joel Rozowsky, Koon-Kiu Yan, Daifeng Wang, Chao Cheng, James B Brown, Carrie A Davis, LaDeana Hillier, Cristina Sisu, Jingyi Jessica Li, et al. Comparative analysis of the transcriptome across distant species. *Nature*, 512(7515):445–448, 2014.

[27] Anthony Gitter, Judith Klein-Seetharaman, Anupam Gupta, and Ziv Bar-Joseph. Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Research*, 39(4):e22–e22, 2010.

[28] Patsy Gómez-Picos and B Frank Eames. On the evolutionary relationship between chondrocytes and osteoblasts. *Frontiers in Genetics*, 6:297, 2015.

[29] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864. ACM, 2016.

[30] Pietro Hiram Guzzi and Tijana Milenković. Survey of local and global biological network alignment: the need to reconcile the two sides of the same coin. *Briefings in Bioinformatics*, page bbw132, 2017.

[31] Somaye Hashemifar and Jinbo Xu. Hubalign: an accurate and efficient method for global alignment of protein–protein interaction networks. *Bioinformatics*, 30(17):i438–i444, 2014.

[32] Wayne B Hayes and Nil Mamano. SANA NetGO: A combinatorial approach to using gene ontology (GO) terms to score network alignments. *Bioinformatics*, 1:8, 2017.

[33] Jialu Hu, Birte Kehr, and Knut Reinert. NetCoffee: a fast and accurate global alignment approach to identify functionally conserved proteins in multiple networks. *Bioinformatics*, 30(4):540–548, 2014.

[34] Ming Huang, ONO Naoaki, Shigehiko Kanaya, Md Altaf-Ul-Amin, et al. BiClusO: A novel biclustering approach and its application to species-VOC relational data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.

[35] Ovidiu D Iancu, Sunita Kawane, Daniel Bottomly, Robert Searles, Robert Hitzemann, and Shannon McWeeney. Utilizing RNA-seq data for de novo coexpression network inference. *Bioinformatics*, 28(12):1592–1597, 2012.

[36] Rashid Ibragimov, Maximilian Malek, Jiong Guo, and Jan Baumbach. Gedevo: an evolutionary graph edit distance algorithm for biological network alignment. In *German Conference on Bioinformatics 2013*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013.

[37] Vinícius Carvalho Jardim, Suzana De Siqueira Santos, Andre Fujita, and Marcos Silveira Buckeridge. BioNetStat: A tool for biological networks differential analysis. *Frontiers in Genetics*, 10:594, 2019.

[38] Zhenhong Jiang, Xiaobao Dong, Zhi-Gang Li, Fei He, and Ziding Zhang. Differential coexpression analysis reveals extensive rewiring of Arabidopsis gene coexpression in response to Pseudomonas syringae infection. *Scientific Reports*, 6, 2016.

[39] Yousang Jo, Sanghyeon Kim, and Doheon Lee. Identification of common coexpression modules based on quantitative network comparison. *BMC Bioinformatics*, 19(8):213, 2018.

[40] Thadeous Kacmarczyk, Peter Waltman, Ashley Bate, Patrick Eichenberger, and Richard Bonneau. Comparative microbial modules resource: generation and visualization of multi-species biclusters. *PLoS Computational Biology*, 7(12):e1002228, 2011.

[41] Maxim Kalaev, Vineet Bafna, and Roded Sharan. Fast and accurate alignment of multiple protein networks. In *Annual International Conference on Research in Computational Molecular Biology*, pages 246–256. Springer, 2008.

[42] Oleksii Kuchaiev, Tijana Milenković, Vesna Memišević, Wayne Hayes, and Nataša Pržulj. Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society Interface*, page rsif20100063, 2010.

[43] Oleksii Kuchaiev and Nataša Pržulj. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, 27(10):1390–1396, 2011.

[44] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, 2008.

[45] Peter Langfelder, Rui Luo, Michael C Oldham, and Steve Horvath. Is my network module preserved and reproducible? *PLoS Computational Biology*, 7(1), 2011.

[46] Luis Guillermo Leal, Camilo López, and Liliana López-Kleine. Construction and comparison of gene co-expression networks based on immunity microarray data from Arabidopsis, Rice, Soybean, Tomato and Cassava. In *Advances in Computational Biology*, pages 13–19. Springer, 2014.

[47] Chung-Shou Liao, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25(12):i253–i258, 2009.

[48] Katya L Mack, Megan Phifer-Rixey, Bettina Harr, and Michael W Nachman. Gene expression networks across multiple tissues are associated with rates of molecular evolution in wild house mice. *Genes*, 10(3):225, 2019.

[49] Noël Malod-Dognin, Kristina Ban, and Nataša Pržulj. Unified alignment of Protein-Protein Interaction networks. *Scientific Reports*, 7(1):953, 2017.

[50] Noël Malod-Dognin and Nataša Pržulj. L-GRAAL: Lagrangian graphlet-based network aligner. *Bioinformatics*, 31(13):2182–2189, 2015.

[51] Hazel N Manners, Ahed Elmsallati, Pietro H Guzzi, Swarup Roy, and Jugal K Kalita. Performing local network alignment by ensembling global aligners. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1316–1323. IEEE, 2017.

[52] Rishi R Masalia, Adam J Bewick, and John M Burke. Connectivity in gene coexpression networks negatively correlates with rates of molecular evolution in flowering plants. *PloS One*, 12(7):e0182289, 2017.

[53] Vesna Memišević and Nataša Pržulj. C-GRAAL: Common-neighbors-based global GRAph ALignment of biological networks. *Integrative Biology*, 4(7):734–743, 2012.

[54] Lei Meng, Joseph Crawford, Aaron Striegel, and Tijana Milenkovic. IGLOO: Integrating global and local biological network alignment. In *12th International Workshop on Mining and Learning with Graphs (MLG) at the 22nd ACM SIGKDD 2016 Conference on Knowledge Discovery & Data Mining (KDD)*, pages 13–17. ACM, 2016.

[55] Lei Meng, Aaron Striegel, and Tijana Milenković. Local versus global biological network alignment. *Bioinformatics*, 32(20):3155–3164, 2016.

[56] Tom Michoel and Bruno Nachtergaele. Alignment and integration of complex networks by hypergraph-based spectral clustering. *Physical Review E*, 86(5):056111, 2012.

[57] M. Milano, P. H. Guzzi, and M. Cannataro. GLAlign: A novel algorithm for local network alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(6):1958–1969, 2019.

[58] Tijana Milenković, Weng Leong Ng, Wayne Hayes, and Nataša Pržulj. Optimal network alignment with graphlet degree vectors. *Cancer Informatics*, 9:CIN–S4744, 2010.

[59] Gianni Monaco, Sipko van Dam, João Luis Casal Novo Ribeiro, Anis Larbi, and João Pedro de Magalhães. A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. *BMC Evolutionary Biology*, 15(1):259, 2015.

[60] Marek Mutwil, Sebastian Klie, Takayuki Tohge, Federico M Giorgi, Olivia Wilkins, Malcolm M Campbell, Alisdair R Fernie, Björn Usadel, Zoran Nikoloski, and Staffan Persson. PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *The Plant Cell*, 23(3):895–910, 2011.

[61] Sergiu Netotea, David Sundell, Nathaniel R Street, and Torgeir R Hvidsten. ComPlEx: conservation and divergence of co-expression networks in A. thaliana, Populus and O. sativa. *BMC Genomics*, 15(1):106, 2014.

[62] Behnam Neyshabur, Ahmadreza Khadem, Somaye Hashemifar, and Seyed Shahriar Arab. NETAL: a new graph-based method for global alignment of protein–protein interaction networks. *Bioinformatics*, 29(13):1654–1662, 2013.

[63] Nam D Nguyen, Ian K Blaby, and Daifeng Wang. ManiNetCluster: A novel manifold learning approach to reveal the functional links between gene networks. *BMC Genomics*, 20(12):1–14, 2019.

[64] Vegard Nygaard, Einar Andreas Rødland, and Eivind Hovig. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, 17(1):29–39, 2016.

[65] Hiroyuki Ogata, Wataru Fujibuchi, Susumu Goto, and Minoru Kanehisa. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Research*, 28(20):4021–4028, 2000.

[66] Michael C Oldham, Steve Horvath, and Daniel H Geschwind. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences*, 103(47):17973–17978, 2006.

[67] Eric N Olson. Gene regulatory networks in the evolution and development of the heart. *Science*, 313(5795):1922–1927, 2006.

[68] Rob Patro and Carl Kingsford. Global network alignment using multiscale spectral signatures. *Bioinformatics*, 28(23):3105–3114, 2012.

[69] Sebastian Proost and Marek Mutwil. PlaNet: Comparative co-expression network analyses for plants. *Plant Genomics Databases: Methods and Protocols*, pages 213–227, 2017.

[70] Pía Francesca Loren Reyes, Tom Michoel, Anagha Joshi, and Guillaume Devailly. Meta-analysis of liver and heart transcriptomic data for functional annotation transfer in mammalian orthologs. *Computational and Structural Biotechnology Journal*, 15:425–432, 2017.

[71] Ashis Saha, Yungil Kim, Ariel DH Gewirtz, Brian Jo, Chuan Gao, Ian C McDowell, Barbara E Engelhardt, Alexis Battle, François Aguet, Kristin G Ardlie, et al. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Research*, 27(11):1843–1858, 2017.

[72] Andrew Schoenrock, Daniel Burnside, Houman Moteshareie, Sylvain Pitre, Mohsen Hooshyar, James R Green, Ashkan Golshani, Frank Dehne, and Alex Wong. Evolution of protein-protein interaction networks in yeast. *PloS One*, 12(3):e0171920, 2017.

[73] Elise AR Serin, Harm Nijveen, Henk WM Hilhorst, and Wilco Ligterink. Learning from co-expression networks: possibilities and challenges. *Frontiers in Plant Science*, 7:444, 2016.

[74] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[75] Yu-Keng Shih and Srinivasan Parthasarathy. Scalable global alignment for multiple biological networks. In *BMC Bioinformatics*, volume 13, page S11. Springer, 2012.

[76] Yong Shui and Young-Rae Cho. Alignment of PPI networks using semantic similarity for conserved protein complex prediction. *IEEE Transactions on Nanobioscience*, 15(4):380–389, 2016.

[77] Bárbara Silva-Vignato, Luiz L Coutinho, Mirele D Poleti, Aline SM Cesar, Cristina T Moncau, Luciana CA Regitano, and Júlio CC Balieiro. Gene co-expression networks associated with carcass traits reveal new pathways for muscle and fat deposition in nelore cattle. *BMC Genomics*, 20(1):32, 2019.

[78] Rohit Singh, Jinbo Xu, and Bonnie Berger. Pairwise global alignment of protein interaction networks by matching neighborhood topology. In *Annual International Conference on Research in Computational Molecular Biology*, pages 16–31. Springer, 2007.

[79] Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, 13(1):328, 2012.

[80] Ruggero Spadafora, Junjie Lu, Radhika S Khetani, Cheng Zhang, Aimee Iberg, Hu Li, Yang Shi, and Paul H Lerou. Lung-resident mesenchymal stromal cells reveal transcriptional dynamics of lung development in preterm infants. *American Journal of Respiratory and Critical Care Medicine*, 198(7):961–964, 2018.

[81] Ralf Steuer, Jürgen Kurths, Carsten O Daub, Janko Weise, and Joachim Selbig. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl_2):S231–S240, 2002.

[82] Joshua M Stuart, Eran Segal, Daphne Koller, and Stuart K Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.

[83] Yihan Sun, Joseph Crawford, Jie Tang, and Tijana Milenković. Simultaneous optimization of both node and edge conservation in network alignment via wave. In *International Workshop on Algorithms in Bioinformatics*, pages 16–39. Springer, 2015.

[84] Vivek Swarup, Flora I Hinz, Jessica E Rexach, Ken-ichi Noguchi, Hiroyoshi Toyoshiba, Akira Oda, Keisuke Hirai, Arjun Sarkar, Nicholas T Seyfried, Chialin Cheng, et al. Identification of evolutionarily conserved gene networks mediating neurodegenerative dementia. *Nature Medicine*, 25(1):152, 2019.

[85] Bruno M Tesson, Rainer Breitling, and Ritsert C Jansen. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics*, 11(1):497, 2010.

[86] Fadi Towfic, Shakti Gupta, Vasant Honavar, and Shankar Subramaniam. B-cell ligand processing pathways detected by large-scale comparative analysis. *Genomics, Proteomics & Bioinformatics*, 10(3):142–152, 2012.

[87] Fadi Towfic, Susan VanderPIas, Casey A OIiver, OIiver Couture, Christopher K TuggIe, M Heather West GreenIee, and Vasant Honavar. Detection of gene orthology from gene co-expression and protein interaction networks. *BMC Bioinformatics*, 11(S3):S7, 2010.

[88] Huynh Thanh Trung, Nguyen Thanh Toan, Tong Van Vinh, Hoang Thanh Dat, Duong Chi Thang, Nguyen Quoc Viet Hung, and Abdul Sattar. A comparative study on network alignment techniques. *Expert Systems with Applications*, 140:112883, 2020.

[89] Panayiotis Tsaparas, Leonardo Mariño-Ramírez, Olivier Bodenreider, Eugene V Koonin, and I King Jordan. Global similarity and local divergence in human and mouse gene co-expression networks. *BMC Evolutionary Biology*, 6(1):70, 2006.

[90] Oren Tzfadia, Tim Diels, Sam De Meyer, Klaas Vandepoele, Asaph Aharoni, and Yves Van de Peer. CoExpNetViz: comparative co-expression networks construction and visualization tool. *Frontiers in Plant Science*, 6:1194, 2016.

[91] Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, page bbw139, 2017.

[92] Vipin Vijayan and Tijana Milenković. Multiple network alignment via multiMAGNA++. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(5):1669–1682, 2017.

[93] Vipin Vijayan, Vikram Saraph, and T Milenković. MAGNA++: Maximizing accuracy in global network alignment via both node and edge conservation. *Bioinformatics*, 31(14):2409–2411, 2015.

[94] Peter Waltman. *Multi-species biclustering: An integrative method to identify functional gene conservation between multiple species*. PhD thesis, Citeseer, 2012.

[95] Daifeng Wang, Eric Pan, Gang Fang, Sunita Kumari, Fei He, Doreen Ware, Sergei Maslov, and Mark Gerstein. Comparative network analysis of gene co-expression networks reveals the conserved and species-specific functions of cell-wall related genes between Arabidopsis and Poplar. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, page 689. ACM, 2013.

[96] Shan Wang, Yanbin Yin, Qin Ma, Xiaojia Tang, Dongyun Hao, and Ying Xu. Genome-scale identification of cell-wall related genes in Arabidopsis based on co-expression network analysis. *BMC Plant Biology*, 12(1):138, 2012.

[97] Michael Watson. CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics*, 7(1):509, 2006.

[98] Claudia C Weber and Laurence D Hurst. Support for multiple classes of local expression clusters in Drosophila melanogaster, but no evidence for gene order conservation. *Genome Biology*, 12(3):R23, 2011.

[99] Koon-Kiu Yan, Daifeng Wang, Joel Rozowsky, Henry Zheng, Chao Cheng, and Mark Gerstein. OrthoClust: an orthology-based network framework for clustering data across multiple species. *Genome Biology*, 15(8):R100, 2014.

[100] Peyman Zarrineh, Ana C Fierro, Aminael Sánchez-Rodríguez, Bart De Moor, Kristof Engelen, and Kathleen Marchal. COMODO: an adaptive coclustering strategy to identify conserved coexpression modules between organisms. *Nucleic Acids Research*, 39(7):e41–e41, 2010.

[101] Peyman Zarrineh, Aminael Sánchez-Rodríguez, Nazanin Hosseinkhan, Zahra Narimani, Kathleen Marchal, and Ali Masoudi-Nejad. Genome-scale co-expression network comparison across Escherichia coli and Salmonella enterica serovar Typhimurium reveals significant conservation at the regulon level of local regulators despite their dissimilar lifestyles. *PloS One*, 9(8):e102871, 2014.

[102] Jiawei Zhang and S Yu Philip. Multiple anonymized social networks alignment. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 599–608. IEEE, 2015.

Chapter 3

pineplot: an R package for visualizing symmetric relationships

This paper introduced a new R package for the visualization of symmetric matrices. Pine plots can be used to visualize large datasets for exploratory data analysis while controlling for different potentially confounding factors. The utility of the package is demonstrated by visualizing gene expression values of tissue-specific genes from RNA-seq data and the clinical factors in a liver disease and a heart disease dataset. This is useful for comparisons made across species as well as other variables as there are many symmetric measures used to represent the relationships between genes (Pearson correlation, Spearman correlation, mutual expression, etc.).

Overall, the package presented in this chapter can be considered applicable to analyzing many types of biological data. As this package applies to many measures used to explore biological datasets, including gene expression, it was used to visualize the statistical test results performed in Chapter 4. Other measures used in this thesis—including Pearson correlation and cosine distance for constructing GCNs and comparing them, respectively—also can utilize this package in future work when visualizing the result of applying these measures.

This paper was accepted as a regular paper at the "The 10th International Conference on Computational Systems-Biology and Bioinformatics (CSBio 2019)".

Citation

Katie L. Ovens, Daniel J. Hogan, Farhad Maleki, Ian McQuillan, and Anthony J. Kusalik. 2019. pineplot: an R package for visualizing symmetric relationships. In Proceedings of the Tenth International Conference on Computational Systems-Biology and Bioinformatics (CSBio 2019). Association for Computing Machinery, New York, NY, USA, Article 6, 1–8. DOI:https://doi.org/10.1145/3365953.3365959

Author contributions

Katie Ovens implemented the pineplot package, designed and implemented case studies, and helped with writing the majority of the paper. Daniel Hogan refactored the code, added regression tests to the pineplot package, and added the heart disease example to the manuscript. Farhad Maleki proposed the idea of a pine plot and helped with writing the first draft of the paper. Ian McQuillan and Anthony Kusalik supervised the work and assisted with revision of the manuscript.

Abstract

An effective publication-quality visualization tells a concise story from data. Methods and tools that facilitate making such visualizations are valuable to the scientific community. In this paper, we introduce *pineplot*, an R package for generating insightful visualizations called pine plots. Pine plots are applicable to a wide variety of datasets and create a holistic picture of the relationship between variables across different experimental conditions. A pine plot provides a means to visualize a group of symmetric matrices, each represented by triangular heat maps. Pine plots can be used to visualize large datasets for exploratory data analysis while controlling for different potentially confounding factors. The utility of the package is demonstrated by visualizing gene expression values of tissue-specific genes from RNA-seq data and the clinical factors in a liver disease and a heart disease dataset. The implementation of *pineplot* offers a straightforward procedure for generating pine plots; full control of the aesthetic elements of generated plots; and the possibility of augmenting generated plots with extra layers of graphical elements to further extend their usability.

3.1 Background

Visualization techniques are indispensable for hypothesis generation, validation of results produced by quantitative data analysis, and presentation of findings. Heat maps are widely used in biological science [17] to visualize data matrices, which could be made from a bivariate similarity/dissimilarity function representing the relationship between two groups of variables. An ordinary heat map can be described as a color-shaded matrix that is constructed based on a data matrix M, where cell (i, j) of the former is assigned a color intensity based on the value of $M_{i,j}$, the component in the i^{th} row and j^{th} column of M [17]. In other words, an ordinary heat map is a means for visualizing a data matrix, where the i^{th} row (j^{th} column) of this matrix corresponds to the variable x_i (y_j), and the value in the i^{th} row and j^{th} column of this matrix is a function of x_i and y_j. Software for generating an instance of such a plot is *heatmap3*, an R package developed by Zhao et al. [18] that was built on the R *heatmap* function and includes features such as customizable legends, side annotations, and colour options.

A typical example of using ordinary heat maps is visualizing gene expression profiles (matrices) where gene expression for a group of genes is measured across different samples—for instance, cancerous versus non-cancerous samples. In a gene expression profile each row corresponds to a gene and each column corresponds to a sample; the (i, j) component of the expression profile represents the measured expression value of the i^{th} gene in the j^{th} sample. A heat map can also be used to visualize some relationship between genes (where the (i, j) represents the relationship between gene i and gene j) or samples (where the (i, j) represents the relationship between sample i and sample j).

A common use-case of heat maps is to visualize the relationship between a group of variables across several experimental conditions. These conditions could be different categories/levels of one or more potentially

confounding factors. Such visualizations can be made using a multipanel figure [3] of several regular heat maps—such as *heatmap3* plots—each illustrating the relationship between variables under study in one experimental condition (for an example, see Figure A.1B).

Most often in data analysis the similarity/dissimilarity function used for the pairwise comparison of variables is a symmetric function, i.e. a bivariate function f for which $f(x_i, x_j) = f(x_j, x_i)$ for any given values x_i and x_j. Examples of such functions are Euclidean distance, Spearman's rank correlation, and Pearson correlation. In these cases, the pairwise comparison of variables across each experimental condition results in a symmetric matrix. Consequently, the resulting heat maps are symmetric and therefore, almost half redundant. While traditional heat maps and those provided by the *heatmap3* package are insightful tools, they do not avoid redundancy as described above. They are also usually limited to the provided options and cannot be easily extended beyond those options. To address these two limitations, we present the *pineplot* R package.

A pine plot serves as an extension to a multipanel figure of ordinary heat maps capable of visualizing the relationships among a group of variables across different experimental conditions using a symmetric similarity/dissimilarity function. It also provides a holistic picture of the relation between variables in a concise and non-redundant manner.

Graphically, a pine plot is a stack of triangular heat maps, where each triangle, hereafter referred to as a layer, depicts a symmetric matrix, e.g. a matrix representing the pairwise relationships between a group of variables in one experimental condition. The sequence of layers represents a third variable in the experiment. When multiple pine plots are placed side-by-side so that the triangular heat maps form a grid, this is referred to as a pine forest. A pine forest allows for the visualization and comparison of a second potentially confounding variable. To include more variables, multiple forests can be used.

In the rest of the paper, we briefly describe the implementation of the *pineplot* R package. Then, the utility of pine plots is first demonstrated using gene expression data for 3 tissues across three species that are extracted from two RNA-seq datasets from works by Merkin et al. [7] and Brawand et al [1]. The resulting pine plots are compared to visualizations using ordinary heat maps. A second demonstration is provided using two disease datasets and utilizes a different symmetric function than the first demonstration. We end the paper with a discussion of additional pine plot applications.

3.2 Implementation

Wilkinson proposed "the grammar of graphics" for describing underlying features governing the composition of all statistical graphics [16]. Based on Wilkinson's grammar of graphics, Wickham developed a layered grammar of graphics and its implementation in R, *ggplot2* [14]. The *ggplot2* package allows generating a statistical graph layer by layer through "a mapping from data to aesthetic attributes (colour, shape, size) of geometric objects (points, lines, bars)" [15]. Therefore, the *pineplot* package was implemented based on

ggplot2.

The developed package offers a straightforward procedure for pine plots that can answer the needs of many users. The package provides functions to build symmetric matrices, visualize these matrices as *ggplot2* triangular heat maps and stack them for easy comparison across group, time, phenotype etc. To create a pine plot, a list of symmetric matrices are required as input as an argument for the *generate_pineplot* function. In addition, optional parameters allow users to manipulate the characteristics of the heat maps that are constructed using the symmetric matrices such as whether to include a legend at the base of the plot. Being implemented on top of *ggplot2*, *pineplot* offers full control of the aesthetic elements of the plot to users familiar with *ggplot2*. The *pineplot* distribution includes examples of adding labels, images, colours, bounding boxes, and legends. Further, the generated pine plots can be augmented with new layers of graphical elements to further extend their usability for special use-cases. Currently, the organization of the features in the pine plots is at the discretion of the user. This allows for flexibility in regards to performing analyses such as clustering prior to generating the pine plot and setting the order of the features.

The *grid* and *gridExtra* R packages are utilized to manipulate the placement of each triangular heat map to create the signature shape of a pine plot. The *grid.arrange* function provided by *gridExtra* also allows users to lay out the pine plots side-by-side for easy comparison as a pine forest.

The development version of the *pineplot* package is available at https://github.com/klovens/pineplot, including instructions for downloading as well as tutorials for each use case we present in the following sections.

3.3 Results

3.3.1 Case study: visualizing tissue-specific genes

The first datasets visualized using pine plots were constructed from RNA-seq expression studies from Brawand et al. [1] and Merkin et al. [7]. These included expression data from three tissues of three different animals: brain, liver, and kidney from each of macaque, mouse, and chicken. This RNA-seq data was read-mapped to the mmul8, mm10, and gal5 genome builds (respectively for the three animals) using STAR 2.5.3. The *quantMode* GeneCounts flag was used in order to obtain the raw counts. These counts were normalized using TMM normalization similar to the meta-analysis performed on both of these datasets by Sudmant et al. [12]. Genes that have been reported as tissue-specific (have gene expression pattern that is specific to a particular tissue) in mouse kidney (*Kl, Pdzk1, Slc12a3, Spp1*, and *Slc34a1*) or liver tissues (*Gnmt, Amdhd1, Ahsg, Ambp, Alb, Slc27a5, Hpx, Apoa1, Fgg*, and *Mat1a*) [11] were selected for visualization. Table A.1 in the Appendix provides the IDs of the RNA-seq samples used for this study.

To visualize the expression status of genes in these samples, we used a measure of mutual expression. For genes g_i and g_j and tissue type t_l, the mutual expression of g_i and g_j in tissue t_l is defined as

$$Mutual\ expression(g_i, g_j) = \frac{\overline{x_{g_i,t_l}} \times \overline{x_{g_j,t_l}}}{\max\limits_{1 \le k \le m_{t_l} \ and \ 1 \le l \le n_t} \overline{x_{g_k,t_l}}^2} \tag{3.1}$$

where $\overline{x_{g_k,t_l}}$ is the average log read count for gene g_k in tissue t_l; n_t is the number of tissues; m_{t_l} is the number of samples of tissue type t_l. Due to the commutative properties of multiplication, mutual expression is a symmetric function. Mutual expression is a number between 0 and 1 inclusive, where 0 is no expression of either gene and 1 is high expression of both genes. Values with intermediate mutual expression could be the result of medium expression of both genes, or higher expression in one gene and lower expression in the other.

The holistic picture provided by the pine plots in Figure 3.1 shows that the genes that were identified by Song et al. [11] as tissue-specific to kidney and liver tended to have moderate mutual expression levels in the kidney and liver gene expression data, respectively. In the pine plot using RNA-seq samples from brain, the majority of these genes had low mutual expression with the exception of *Apoa1* in chicken. However, several genes predicted as being tissue-specific are also expressed in different tissue types. *Spp1*, for example, is a multifunctional gene known to be expressed in many tissues including kidney, liver, brain, bone, and dentin [10, 13].

As depicted by the pine plots in Figure 3.1, the mutual expression results show that in liver tissue, *Spp1* is expressed intermediately compared to the other genes considered "kidney-specific". The pine plots also show that *Pdzk1*, predicted as being "kidney-specific" by Song et al. [11], was expressed intermediately in the liver tissue. This gene is known to bind to and mediate the localization of cell surface proteins. It has also been reported elsewhere as active in mice livers [4]. Furthermore, the pine plot layer visualizing gene expression in the chicken kidney samples shows moderate mutual gene expression for many of the genes predicted as "liver-specific".

Figure 3.1 also makes it visually apparent that the same tissue shows similar expression patterns within all three species. This supports the widely accepted idea that gene expression is more similar in the same tissue in different species compared to different tissues within the same species [12]. By visualizing the expression of genes of interest in this manner, insight can be derived about the relationships between the expression of particular genes, and if these relationships are conserved. By analyzing the differences between the mutual expression of g_i with g_j, either between multiple tissues or multiple organisms, it is possible to assess how gene expression dependencies are different between the tissues or changed throughout evolution between organisms. Pine plots allow for these differences to be visually assessed, either between two specific genes, or pairwise genes more broadly. This makes pine plots useful for exploratory analyses.

The similarity function can also be changed depending on what kind of relationship a researcher wishes to visualize. As such, we provide a second case study utilizing correlation as the symmetric function.

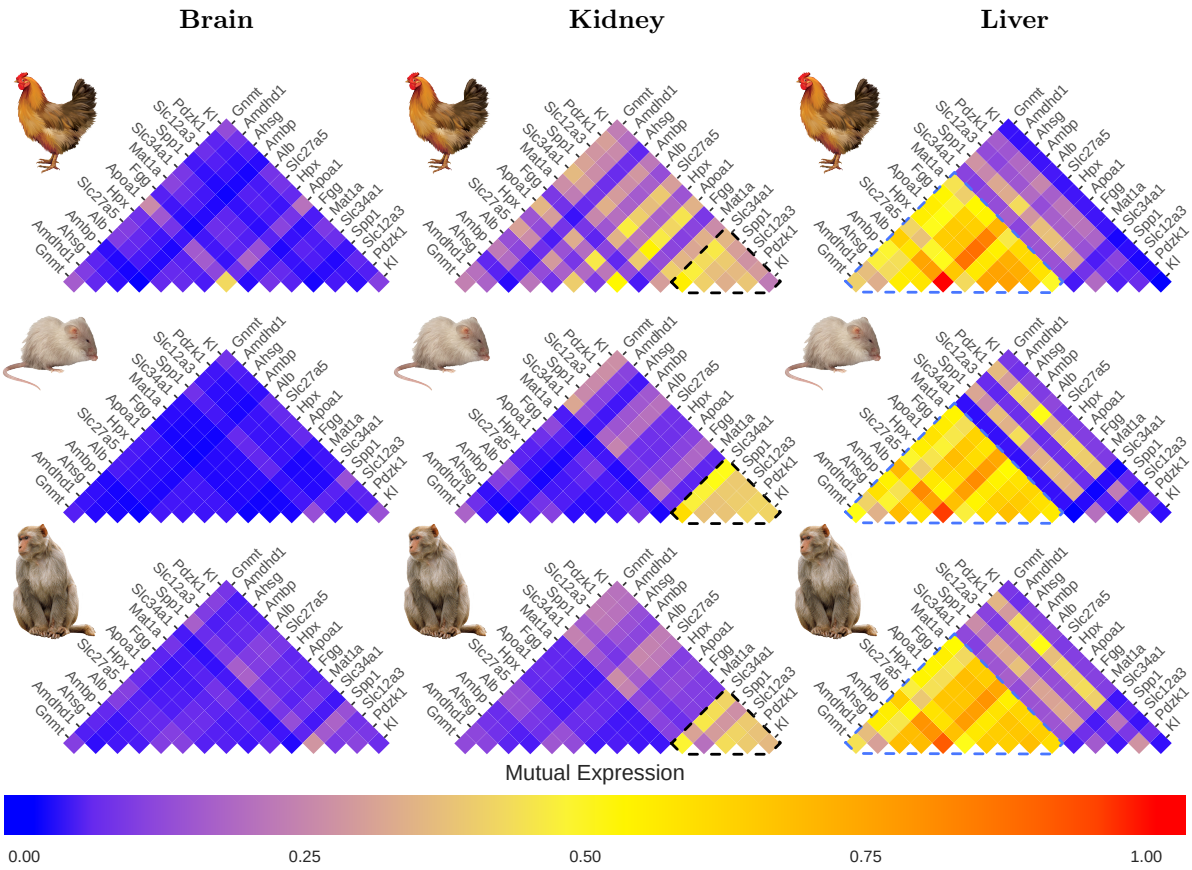


Figure 3.1: Example of a pine forest of kidney and liver-specific genes in three tissues—brain, kidney, and liver—across three species—macaque, mouse, and chicken. Pine plots in the left, middle and right illustrate the mutual expression in brain, kidney, and liver, respectively. In each pine plot, the bottom, middle, and top layers correspond to macaque, mouse, and chicken, respectively. The relationship between the expression of these genes is measured in terms of mutual expression (see Equation 3.1). Note that the order of these genes is arbitrary, yet consistent, across all of the pine plots.. The colour blue indicates both genes are likely not expressed or expressed at very low levels. Yellow indicates that both genes could be expressed at intermediate levels or one gene could be expressed highly and the other scarcely. Red indicates that both genes are highly expressed. The dotted triangles indicate the genes that are predicted in Song et al. [11] as tissue-specific genes, showing the package has the ability to easily highlight areas of interest in the plots. Genes predicted as liver-specific genes are highlighted with blue dashed lines in the plots containing liver tissue samples. Genes predicted as kidney-specific genes are highlighted with black dashed lines in the plots containing kidney tissue samples. In all three species, brain shows very little expression for any of the kidney and liver-specific genes with the exception of *Apoa1* in chicken. In all three species, the liver-specific genes are highly mutually expressed in the liver tissue samples. The kidney-specific genes are expressed at lower levels in the liver tissue samples, but *Spp1* and *Pdzk1* are more intermediately mutually expressed with most of the other genes. This observation suggests that these genes possibly are not specific to kidney. Considering the small sample size, these observations deserve further investigation.

3.3.2 Case study: disease datasets

The two datasets visualized for this case study contained clinical variables originating from a heart disease dataset and a liver disease dataset.

The Cleveland heart disease dataset

The Cleveland heart disease dataset contains 14 attributes across 303 subjects. The attributes were selected for their relevance to heart disease and include variables such as age, sex, cholesterol level, and maximum heart rate achieved.

Pine plots were used to examine the relationship between attributes relevant to heart disease when controlling for variables of sex and age (see Figure 3.2). For the age attribute, 3 levels were defined (i.e. age groups) using the *cut* R function. Pearson correlation was measured between 5 clinical measures across subjects within each category defined for age and sex. Correlation was calculated as shown in Equation 3.2 after centering and scaling the variables based on mean and standard deviation estimators. For a clinical measure X and another clinical measure Y, correlation is calculated as follows.

$$r(X_{age,sex}, Y_{age,sex}) = \frac{cov(X_{age,sex}, Y_{age,sex})}{\sigma_{X_{age,sex}} \sigma_{Y_{age,sex}}} \tag{3.2}$$

where *cov* is covariance and σ is the standard deviation.

In both the male and female pine plots, there is evidence of different relationships between variables for one sex versus the other. For example, females in the 29 to 45 age bracket have lower correlation, and possibly inverse correlation, between serum cholesterol, maximum heart rate achieved, and resting electrocardiograph results compared to males in the same age bracket. Furthermore, fasting blood sugar also has low correlation with maximum heart rate achieved and resting electrocardiograph results. This suggests that the relationship between these attributes is weakened possibly by a change in cholesterol levels in the female group in this age bracket. Also, both cholesterol and fasting blood sugar are highly correlated with each other. However, when interpreting the male samples a different pattern is observed, suggesting that these two factors behave differently in this particular group. It is possible that more samples from females are required in order to detect the relationship observed in the male samples or vice versa. However, it could also suggest that the underlying causes of the heart disease in the females studied versus the males are different, possibly with different environmental or genetic influences. If one were to combine the male and female results in this scenario, it is possible that this pattern could be hidden or cancelled out by the values of these variables in the male group or vice versa. These interpretations could be explored further with the addition of more samples to each group or with a new/different cohort with similar demographic characteristics to determine if the pattern remains consistent or becomes more apparent.

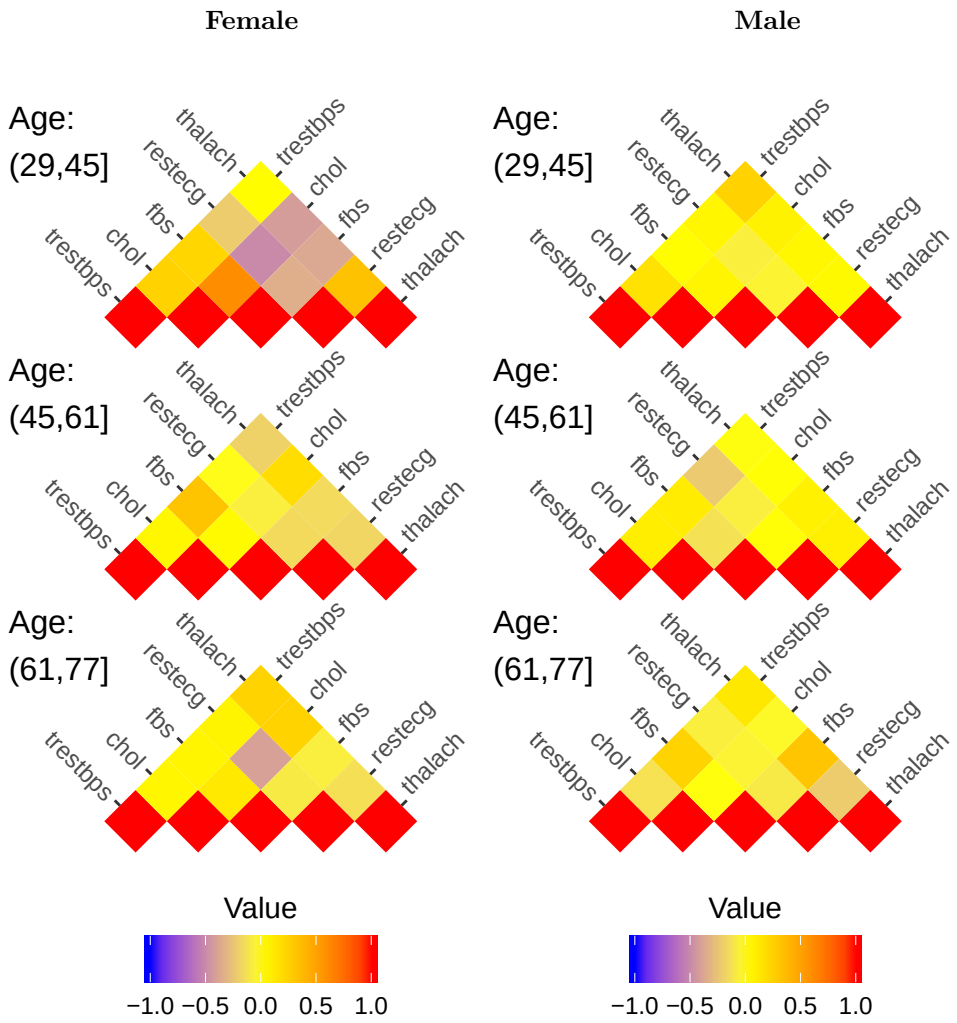


Figure 3.2: Example of two pine plots (pine forest) visualizing the correlation between 5 clinical measures (thalach=maximum heart rate achieved, restecg=resting electrocardiographic results, fbs=fasting blood sugar, chol=serum cholestoral, trestbps=resting blood pressure) while controlling for confounding factors of sex and age. The pine plots for females are on the left and for males are on the right.

Liver disease dataset

A liver disease dataset is utilized in a third demonstration. The dataset was from the University of California at Irvine (UCI) Machine Learning Repository [2] and contained 7 clinical variables potentially relevant to liver disease across 583 patients. A description of these clinical variables is provided in Table A.2 in the Appendix. Again, Pearson correlation was used as a measure of similarity between these variables.

Authors combining this data with other liver patient datasets discovered a significant difference in alkaline phosphatase (Alkphos), aspartate aminotransferase (Sgot), and alanine aminotransferase (Sgpt) between liver disease and healthy control samples [8]. The pine forests for healthy and disease groups shown in Figures 3.3 and 3.4 appear to support visually the importance of Sgot and Sgpt, which was observed using statistical tests in the original publication. Sgot and Sgpt have a much stronger positive correlation in the disease class than in the healthy class for the entire population in each group visualized in Figure 3.4. This indicates that a coordinated change in the levels of these two enzymes could be highly informative when diagnosing liver disease, regardless of sex or age. However, Alkophos is not as consistently correlated with these two enzymes and its relationship with the other clinical variables changes depending on the age group as well as sex. However, females younger than 45 years old with liver disease have a strong positive correlation observed in Alkphos, Sgot, and Sgpt. This is also the case in older males without liver disease. In the healthy liver group for females, depending on the age, the correlation between these variables is either inconsistent across ages or non-existent.

To understand the relationship between each attribute and how these relationships change due to the influence of confounding factors is obviously important. The liver disease example uses multiple pine forests for comparison of three variables. Such an analysis could be useful for all manner of large datasets with large numbers of variables.

3.4 Discussion

The utility of the *pineplot* package was demonstrated using three considerably different case studies utilizing different datasets and symmetric functions to visualize the relationships between variables. Pine plots have also been used in other contexts, demonstrating their flexibility for visualizing data in many scenarios.

Pine plots have been used in the context of gene set analysis [5, 6]. These plots were used in the systematic study of how sample size influences downstream data analysis of gene set enrichment analysis methods, controlling for the confounding variable of sample size [5]. This visualization technique was also utilized to compare the results across gene set analysis methods where the potentially confounding factors were dataset and number of gene sets reported as significant [6]. The Jaccard index was the similarity measure utilized in both cases.

Two alternatives to visualizing the data in the case studies presented in Section 6.3 are to generate stacks

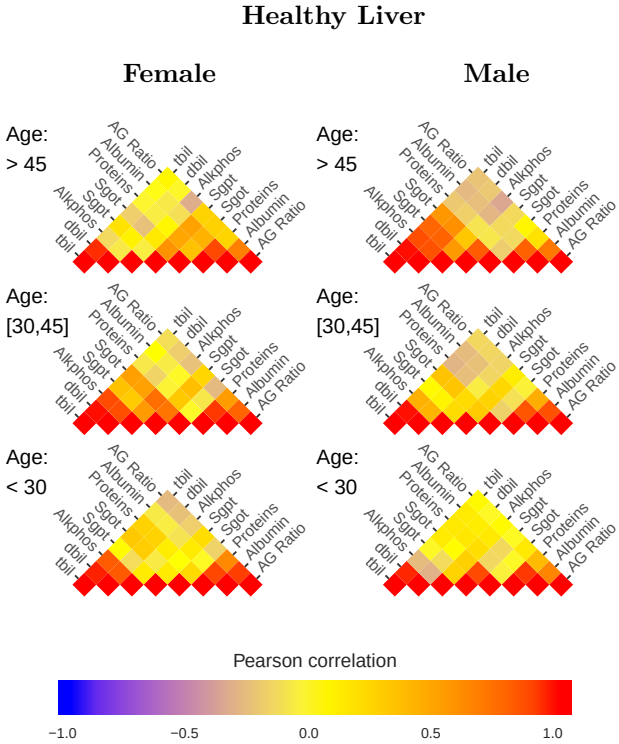


Figure 3.3: Example of a pine forest of 7 clinical variables from healthy males (right) and females (left) with no liver disease separated into age groups. The red cells in the plot indicate positive correlation between variables, while blue cells indicate negative correlation between variables. Yellow indicates little or no correlation between the variables.

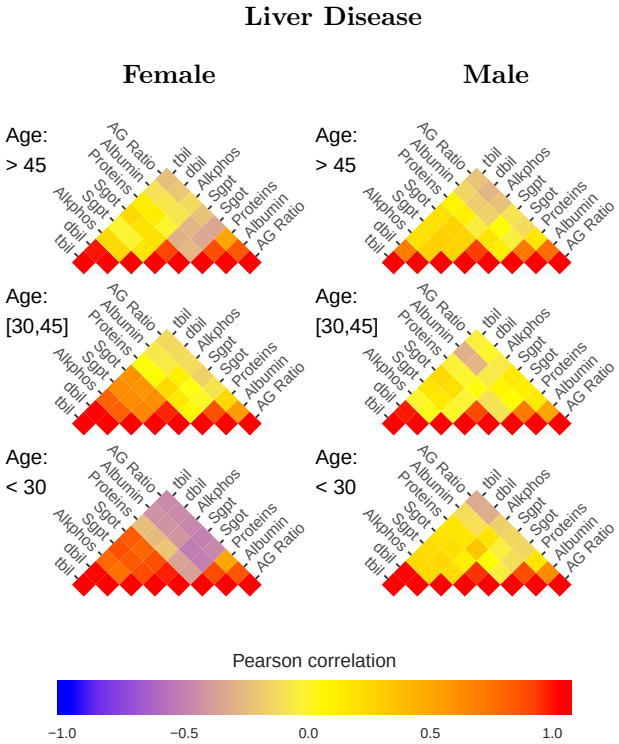


Figure 3.4: Example of a pine forest of 7 clinical variables from males (right) and females (left) with a liver disease separated into age groups. The red cells in the plot indicate positive correlation between variables, while blue cells indicate negative correlation between variables. Yellow indicates little or no correlation between the variables.

of full heat maps or another multipanel figure such as a scatter plot array. Examples of these alternative visualizations are provided in the Appendix.

The first alternative is to retain the full heat maps and stack them in the same manner as a pine plot. However, as shown in Figure A.1, the redundant information makes these figures much larger. Consequently, this can make it more challenging to visualize a larger number of variables. The information in the second half of the heat map is redundant and serves as unnecessary distraction when interpreting a visualization. The removal of the redundancy using the *pineplot* package reduces the size of the plots. This is important when trying to keep a manuscript within page count limits. It also means that it is less likely that the figure needs to be excessively reduced in size to fit within column or page boundaries. The latter can sometimes lead to detail that is too small to be legible.

The second alternative of a scatter plot array is shown in Figure A.2 for the liver disease dataset. In the scatter plots colour and shape of the points represent the sex as well as different age groups. The scatter plot array is not as effective as the pine plots in Figures 3.2, 3.3, and 3.4. It is possible to observe that some clinical variables have stronger linear relationships than others, such as dbil and tbil, but any further interpretation of the characteristics of the points in each scatter plot is challenging without having very large images. One suggestion is to split the groups over multiple scatter plot arrays. However, this takes up a significant amount of space in comparison, since the plots also need to be large enough to see any of the points within a single cell of the scatter plot array. It is also not as obvious what the relationship is between the variables when comparing the same variables across different scatter plot arrays. Although these scatter plots could be rotated to the same shape and configuration as a pine plot, it is not an intuitive way to interpret them.

In data analysis, it is typical that one or more variables act as confounding factors. These variables, although not being considered by the research hypothesis, may affect the other variables under study. Ignoring confounding variable(s) may lead to erroneous conclusions [9]. Ideally, confounding variables are known *a priori* and are accounted for in experiment design; however, it is often not the case due to incomplete domain knowledge, dynamic nature of data, or the large number of potentially confounding factors. A full factorial design is an attempt to address the known confounding variables. Pine plots can be used in this case to visualize the relationship between variables across different levels (values) of each confounding factor. Pine plots can also be used to check the interaction between variables for detecting confounding factors that have not been identified in the experimental design. Due to the human capacity to interpret visual data as compared to quantitative measures, a technique capable of visualizing confounding factors is a valuable tool for data analysis.

Visualization techniques such as pine plots can be utilized for gene expression studies that include the observation of differences in correlation of gene expression, while controlling for variables that tend to cause batch effects, such as sex, species, extraction protocol etc. Also, pine plots could be used for exploratory data analysis using large datasets typically used for machine learning or deep learning projects.

Future additions to the *pineplot* package could also include compatibility with animation packages that also extend the grammar of graphics of *ggplot2*. This could be useful for visualizing time-dependent datasets. Additional benefits of making the package compatible with animations include saving further space by combining figures that would otherwise be presented side-by-side as well as potentially utilizing the animations for improved visualization for presentations or online content.

3.5 Conclusion

In this paper, we introduced the *pineplot* package for constructing a stack or grid of triangular heat maps for visualizing relationships among a group of variables across different conditions, where the base heat maps use a symmetric similarity/dissimilarity function. Pine plots avoid uninformative redundancy and produce concise visualizations.

The implemented *pineplot* package offers a straightforward procedure for generating pine plots; full control of the aesthetic elements of generated plots; and the possibility of augmenting generated plots with extra layers of graphical elements to further extend its usability. The concise representation of a pine plot facilitates creating a holistic picture of the relationship between variables across different experimental conditions. In addition, pine plots are an insightful visualization tool for a wide variety of datasets.

References

[1] David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, et al. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343, 2011.

[2] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

[3] Johannes Graumann and Richard Cotton. multipanelfigure: Simple assembly of multiple plots and images into a compound figure. *Journal of Statistical Software*, 84(1):1–10, 2018.

[4] Olivier Kocher, Ayce Yesilaltay, Ching-Hung Shen, Songwen Zhang, Kathleen Daniels, Rinku Pal, Jianzhu Chen, and Monty Krieger. Influence of pdzk1 on lipoprotein metabolism and atherosclerosis. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1782(5):310–316, 2008.

[5] Farhad Maleki, Katie Ovens, Ian McQuillan, and Anthony J Kusalik. Sample size and reproducibility of gene set analysis. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 122–129. IEEE, 2018.

[6] Farhad Maleki, Katie L. Ovens, Elham Rezaei, Alan M. Rosenberg, and Anthony J. Kusalik. Method choice in gene set analysis has important consequences for analysis outcome. In *10th International Conference on Bioinformatics Models, Methods, and Algorithms*, pages 43–54, Prague, Czech Republic, February 2019.

[7] Jason Merkin, Caitlin Russell, Ping Chen, and Christopher B Burge. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–1599, 2012.

[8] Bendi Venkata Ramana, M Surendra Prasad Babu, and NB Venkateswarlu. A critical comparative study of liver patients from USA and India: an exploratory analysis. *International Journal of Computer Science Issues (IJCSI)*, 9(3):506, 2012.

[9] James J Schlesselman. Assessing effects of confounding variables. *American Journal of Epidemiology*, 108(1):3–8, 1978.

[10] J Sodek, B Ganss, and MD McKee. Osteopontin. *Critical Reviews in Oral Biology & Medicine*, 11(3):279–303, 2000.

[11] Yan Song, Jinsoo Ahn, Yeunsu Suh, Michael E Davis, and Kichoon Lee. Identification of novel tissue-specific genes by analysis of microarray databases: a human and mouse model. *PloS One*, 8(5):e64483, 2013.

[12] Peter H Sudmant, Maria S Alexis, and Christopher B Burge. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biology*, 16(1):287, 2015.

[13] Shunyan Weng, Liang Zhou, Lei Han, and Yunsheng Yuan. Expression and purification of non-tagged recombinant mouse SPP1 in E. coli and its biological significance. *Bioengineered*, 5(6):405–408, 2014.

[14] Hadley Wickham. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 19(1):3–28, 2010.

[15] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, 2016.

[16] Leland Wilkinson. The grammar of graphics. In *Handbook of Computational Statistics*, pages 375–414. Springer, 2012.

[17] Leland Wilkinson and Michael Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.

[18] Shilin Zhao, Yan Guo, Quanhu Sheng, and Yu Shyr. Advanced heat map and clustering analysis using heatmap3. *BioMed research international*, 2014, 2014.

Chapter 4

The impact of sample size on the reproducibility of gene co-expression networks

Since no gold standard co-expression dataset is available to evaluate the influence of sample size on accuracy of co-expression network prediction, a similar methodology to one utilized by Maleki et al. [19] is used to evaluate the reproducibility of gene set analysis methods to determine how methods behave using real gene expression data as opposed to synthesized datasets. This paper explores the consistency of co-expression networks when constructing these networks using distinct gene expression samples originating from real RNA-seq datasets. The paper also provides a methodology to determine the number of samples to produce stable networks as it can fluctuate depending on how the networks are constructed.

This chapter demonstrates some of the limitations discussed in Chapters 2 and 7 when constructing and analysing GCNs constructed using real datasets and why the results of many of the studies done with GCNs with small sample sizes should be taken with a grain of salt. The paper offers insight into strategies and quantitative measures that can be used to compare GCNs within the same tissue, condition, or species. These methods can be used to investigate how comparable different GCNs are based on the reproducibility of GCN construction using a particular dataset.

This paper was accepted at the 11th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB).

Citation

K. Ovens, B. Frank Eames, and I. McQuillan. 2020. In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, September 21-24, 2020, Virtual Event, USA. ACM, New York, NY, USA, 10 pages. DOI:https://doi.org/10.1145/3388440.3412481

Author contributions

Katie Ovens adapted the methodology for assessing the effect of sample size on gene set analysis methods to make it applicable to evaluating large and fully connected gene co-expression networks, implemented the pipeline for generating gene co-expression networks, performed the statistical analysis and data visualization, and wrote the paper. Ian McQuillan and Brian Eames supervised the research and helped edit and revise the paper.

Abstract

Identifying relationships between genes facilitates the comparison of different cell types at the transcriptomic level. Gene expression data such as RNA-seq can be used to construct co-expression networks, which is one means in systems biology to describe the coordinated expression patterns among genes across samples. Currently, there is no consensus as to the number of samples required to construct a reproducible gene co-expression network. Indeed, irreproducibility of gene expression experiments is a major challenge, and small sample sizes tend to be one of the major causes. However, recommending a single sample size that applies to all scenarios may not be practical. As such, we utilize a systematic, quantitative approach to study the effect of sample size on the reproducibility of constructing large, fully-connected gene co-expression networks using several correlation-based measures or mutual information. This approach does not require synthetic datasets that are constructed based on oversimplified assumptions nor is it dependent on known functional annotations. Further, we describe two similarity measures to measure consistency and use them to determine if the biological variance present within samples impacts the rate at which the networks will stabilize and compare to networks with randomly reassigned nodes. Our results show that the required number of samples to construct consistent co-expression networks could be influenced by the tissue type used to construct the networks as well as the similarity measure used to measure consistency.

4.1 Introduction

The availability of high-throughput transcriptomic datasets allows for the investigation of the coordinated patterns in gene expression data using gene co-expression networks (GCNs). GCNs can be represented as an undirected graph, where the nodes of the graph represent genes, and the edges are weighted to represent the strength of the transcriptional relationship between genes. Due to the availability of public gene expression datasets for model and non-model organisms, GCNs have been widely used to study co-regulation of genes across phenotypes [25]. Correlation-based methods—such as WGCNA [15]—as well as mutual information-based methods [24] are commonly used for measuring the relationship between genes and detecting co-regulation. Possible reasons for co-regulation between genes include that they are active in the same pathway, or they share a common biological function, location or process [13]. As such, co-expression networks have been utilized for functional gene annotation, to identify disease driver genes, and they can be an important step towards identifying regulatory genes [26]. Understanding these networks is also valuable for understanding development and they can provide evidence for differences between cells or tissues [12, 17, 21].

The majority of typical transcriptome datasets tend to be small in terms of the number of samples. Using small sample sizes for building GCNs may lead to variable or erroneous results [16]. Moreover, using large samples sizes often is not possible due to cost and time constraints, and also limited resources. This poses the question about the optimal number of samples to be used for building GCNs. Currently, there is no

consensus among researchers regarding the number of samples necessary to form an acceptably reproducible co-expression network [2, 3, 5, 11, 16]. This lack of consensus could be due to the heterogeneity of data from different phenotypes [10]. Also, these guidelines have often been developed using simulated data based on oversimplifying assumptions or relying on functional annotations that are not equally represented across phenotypes [20].

Due to the lack of gold standard datasets, i.e. data sets with known interaction patterns between all genes, evaluating the accuracy of GCNs is a challenging task [9]. Simulated datasets, with known network structures, have been commonly used for evaluation of GCNs [2]. These simulated datasets are often built with a small number of genes that are substantially different from real-world GCNs [1, 3]. Also, these networks have often been designed with normally distributed edge weights that are unlikely to represent real-world applications [2]. Geier et al. [11], using linear Gaussian dynamic Bayesian network and discrete dynamic Bayesian network methods, reported that at least 20 samples are required to outperform a random prediction with observational noise levels of 20%. Altay et al. [3], based on a simulation study, reported that an estimated 64 samples are enough for obtaining the best possible predictions when considering precision as the performance measure. The results of these simulation studies might not generalize to different methods or phenotypes of interest. Also, these evaluations often consider microarray datasets with a relatively simple organism, which means the study may not apply to highly complex organisms, such as vertebrates.

To study the reproducibility of GCNs across several expression measurement platforms, Vinciotti et al. [27] built GCNs from total blood samples in human using several platforms. Using sparse Gaussian graphical models, they showed that the generated GCNs are inconsistent across platforms. This suggests that the guideline for the number of samples to be used for a given experiment using a specific platform might not be generalizable to other platforms.

GCNs have also been evaluated based on the functional information captured by the networks. Ballouz et al. [5] compared GCNs based on correlations of node degree as well as the functional information as encoded by Gene Ontology (GO), KEGG, and Reactome databases. Based on a guilt by association approach, they developed a machine learning model to determine if neighbouring genes have the same functional annotation based on the databases mentioned above. Their results suggested that 20 samples should be sufficient to construct GCNs. Liesecke et al. [16] used a down-sampling approach [14] as well as a similar methodology described by Ballouz et al. to determine how well GCNs built with subsets of samples capture the same information as GCNs constructed using all samples of a dataset. They reported that a combination of more than 100 samples could generate reproducible networks. While relying on known functional connectivity may be applicable to evaluating GCNs for well-characterized organisms and conditions, it is not reliable for the study of less characterized organisms and phenotypes, where known functional annotations are not available or incomplete.

The reproducibility of GCNs depends on many factors including the heterogeneity of the data, the phenotype of interest, the high-throughput technology being used, etc. Therefore, specifying a single number as an

appropriate sample size for building GCNs might not be a generalizable approach. This could be the reason behind the lack of consensus for the required samples size for building reproducible GCNs. Also, methods utilizing simulated data have been reported to lead to false conclusions by ignoring the characteristics of real data [18]. Furthermore, relying on known functional components for evaluating GCNs limits the utility of the evaluation methods to only well-studied phenotypes. In this paper, we present an approach for estimating the required sample size for a given phenotype of interest using real data, and independent of the availability of known functional annotations. The proposed method is applicable across different methods, phenotypes, platform, and provides a reliable estimate of sample size by capturing the true characteristics of real data.

The rest of the paper is organized as follows. We describe the procedure and similarity scores used for measuring consistency between co-expression networks constructed with different sample sizes in Section 4.2. The results of these comparisons are presented in Section 4.3 and the discussion of these results is in Section 4.4. Finally, we conclude the paper in Section 4.5. We also make the code to compare networks, including the tissue-specific genes selected for co-expression network construction publicly available [22].

4.2 Methods

The goal of this section is to describe a methodology for assessing consistency of co-expression networks without using artificial datasets. To accomplish this, we first describe a procedure for generating co-expression networks using different subsets of samples from a larger gene expression dataset. Furthermore, we describe two measures utilized in order to make comparisons between co-expression networks which are used to measure their consistency.

4.2.1 Data

RNA-seq data from brain (2642 samples), heart (861 samples), and muscle (803 samples) tissues were obtained from the GTEx consortium, which provides a view of the human transcriptome across dozens of tissues. The dataset is based on a large-scale RNA-Seq experiment of postmortem tissue from hundreds of human donors [7]. The gene expression counts were obtained from the GTEx Portal on 17/01/2020.

A visualization of the network consistency evaluation procedure is shown in Figure 4.1 and it is described in detail in the following sections.

4.2.2 Network construction

A co-expression network can be represented as a weighted (undirected) graph G that consists of a set of nodes and edges. The number of genes remains consistent for all the co-expression networks constructed in this study. A group of 10,000 randomly selected genes (to account for memory constraints) from the original dataset was used for constructing all the co-expression networks for brain, heart, and skeletal muscle. This

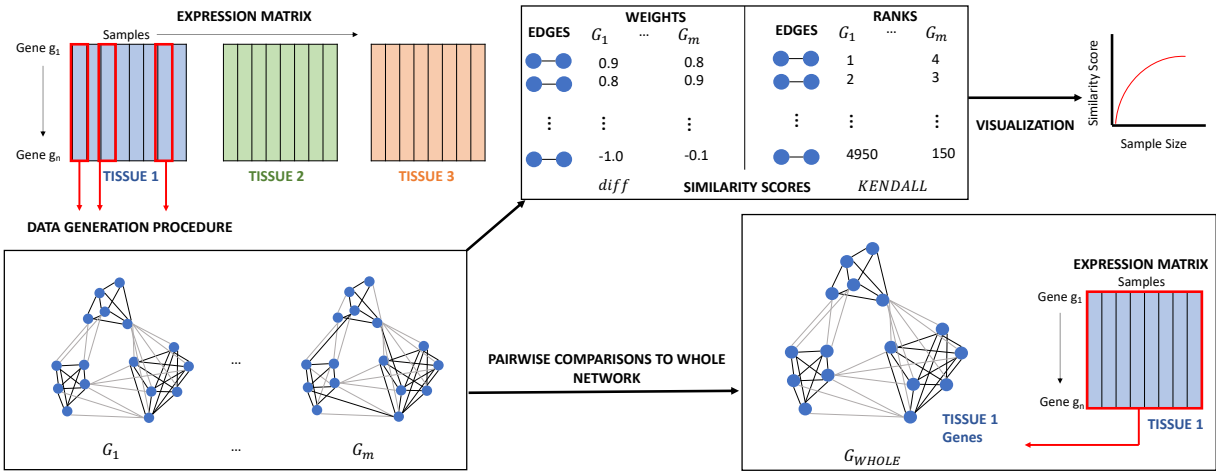


Figure 4.1: Methodology used to compare gene co-expression networks constructed using different sample sizes. In this study, samples from 3 tissues are selected for the experiments. This diagram visualizes the procedure using one tissue, labelled "TISSUE 1". The genes $g_1, ..., g_n$ are genes from the samples of interest. During the dataset generation procedure, we randomly select s number of samples from one tissue. We do this m times (without repeatedly using any sample) and generate a gene co-expression network created from each subset of s samples. Each subset of samples generated using samples from "TISSUE 1" is used to generate a separate co-expression network. A total of m networks are compared using statistical tests to compare the difference in edge weights between networks (absolute difference similarity score) and to compare the order of edges in each network (Kendall concordance coefficient test). The results of these tests are then plotted for each sample size. The networks are also compared to a network constructed using all samples of "TISSUE 1".

resulted in a total of $E = 49,995,000$ edges to make each fully connected network. Each edge between a pair of genes is weighted based on a measure of correlation.

Furthermore, four measures of the relationships between genes are used to construct the co-expression networks (the weights on the edges) to determine if the results are similar for each of the measures of correlation used. The four measures tested were (1) Pearson correlation, (2) Spearman correlation, (3) WGCNA (signed, $\beta = 12$), and (4) mutual information.

4.2.3 Network comparison

Since no gold standard co-expression dataset is available to evaluate the influence of sample size on accuracy of co-expression network prediction, a similar methodology to one utilized by Maleki et al. [19] is used to evaluate the reproducibility of gene set analysis methods. Given an original dataset D with S_T samples for a particular tissue T, where S_T should be relatively large (> 50), we randomly sample (without replacement) s samples (out of all S_T samples where $s << S_T$). For each sample size s where $3 \leq s \leq 50$, the dataset generation procedure from [19] is used to assemble m replicate datasets $D_1, \ldots, D_m$. A maximum sample size of 50 was selected to ensure none of the replicate datasets generated from a tissue would contain any overlapping samples. For all experiments performed using each tissue, we construct 10 datasets ($m = 10$), which are used to construct their corresponding co-expression networks $G_1^s, \ldots, G_m^s$.

Two different measures were used to measure the similarity between each replicate network. Since the nodes within each network of a single tissue remain the same regardless of the sample subset analysed, the networks were compared based on differences between their edge weights. The networks were compared using both a normalized absolute difference between edge weights shown in Equation 4.1, and using Kendall correlation coefficient [4], explained next.

After constructing m co-expression networks $G_1^s, \ldots, G_m^s$ using the expression data for genes across s samples, the weight of the edge between node i and j in G_p^s is denoted as $W_{ij}^{G_p^s}$, where $1 \leq p \leq m$. The normalized absolute difference between two co-expression networks is calculated as follows.

$$diff(G_p^s, G_q^s) = \frac{\sum_{i=1}^{l} \sum_{j=i+1}^{l} |W_{ij}^{G_p^s} - W_{ij}^{G_q^s}|}{2E} \tag{4.1}$$

Assuming $-1 \leq W_{ij}^{G_p^s} \leq 1$, as is the case with measures such as Pearson correlation, the maximum absolute difference between two edge weights can be 2. Therefore, we normalize by $2E$ (recall that E is the number of edges, which is the number of additions in the numerator of Equation 4.1, and therefore the numerator ranges from 0 to $2E$). Then the normalized value is between 0 and 1. To measure the similarity between two graphs, we use $sim(G_p^s, G_q^s) = 1 - diff(G_p^s, G_q^s)$. A similarity score of 0 indicates no similarity between the weights in the two networks predicted and a similarity score of 1 indicates the networks have the same weights on their corresponding edges. Although it is highly unlikely that a score of 1 would ever be achieved in networks constructed using real biological data due to factors such as biological variation, we hypothesize that for a large enough sample size, similarity between networks that use replicate datasets should ideally be 1. In order to determine if the absolute difference in edge weights using different sample sizes is significant, a Kruskal-Wallis test is performed followed by a Dunn post-hoc analysis. The p-values are adjusted for multiple comparisons using the Benjamini-Hochberg procedure [6].

The second similarity measure utilized is the Kendall concordance coefficient test (W) [4] that is used to determine the association between the replicate datasets of the same sample size. This score is indicative of whether the *order* of the edges when sorted by their edge weights is conserved across replicate networks. Using this test, a value of 0 indicates no agreement between the order of the edges and a value of 1 indicates complete agreement. In co-expression networks, if the edges are ranked according to their weights, the order should have some correspondence between networks constructed using samples from the same tissue. The Kendall W scores for each tissue type are also calculated after permuting the nodes used to construct each co-expression network to determine the improvement in this score as sample size increases that is expected by chance.

In this way, we are able to quantify the consistency between the networks constructed using smaller subsets of a larger gene expression dataset, and we can determine to what degree it is beneficial to increase the number of samples to construct co-expression networks. The code to compare networks is publicly available as a Github repository [22].

4.3 Results

The co-expression networks generated for a particular sample size (and tissue) were compared in a pairwise fashion using both similarity score, and the order of edges sorted by edge weight. Figure 4.2 shows the results of the Kendall concordance test across three tissues when utilizing different correlation measures to construct the networks. These plots show an increase in Kendall W scores as sample size increases regardless of the tissue type or the method used to construct the networks. Also, the networks constructed using samples that originated from brain tissue consistently have a higher Kendal W score compared to heart and skeletal muscle samples. Networks generated from heart samples also consistently had higher scores than networks constructed using skeletal muscle.

Figure B.4 in the Appendix illustrates the results of the permutation tests conducted—by random re-assignment of nodes before the construction of each replicate network. This was done to ensure that the observed increase in Kendall W scores was not based on irrelevant network characteristics. These results showed that in a permuted scenario, there is no significant difference between reproducibility in the tissue types. Also, there was no increase in the score as sample size increased.

Figure B.3 illustrates the change in similarity scores when utilizing different numbers of samples to construct the networks. For the sake of visualization, the scores are only shown above 0.5 as no similarity score fell below this value. As with the Kendall W scores, the similarity scores also increased as sample size increased. However, the similarity did not increase substantially when more than 10 samples were used to construct the networks. This was observed for all three tissues, where similarity scores ultimately plateau around 0.9. Figure B.3 shows the results when using Spearman correlation to construct the networks. Pearson correlation, WGCNA, and mutual information also showed a similar pattern with WGCNA being the most consistent of these methods.

Table B.1 in the Appendix presents the results of the Kruskal-Wallis test, which showed a significant difference between the results of the similarity scores calculated for different sample sizes. This trend was observed across all three tissues and methods of constructing the co-expression networks. We conducted the post-hoc analysis using the Dunn test. Figure 4.4 illustrates the pine plot [23] of the Dunn post-hoc analysis results. It was observed that there were significant differences in the similarity scores for small sample sizes versus large sample sizes. Further, the difference between scores achieved using samples sized more than a specific threshold is not significant. However, this threshold is tissue and method dependent.

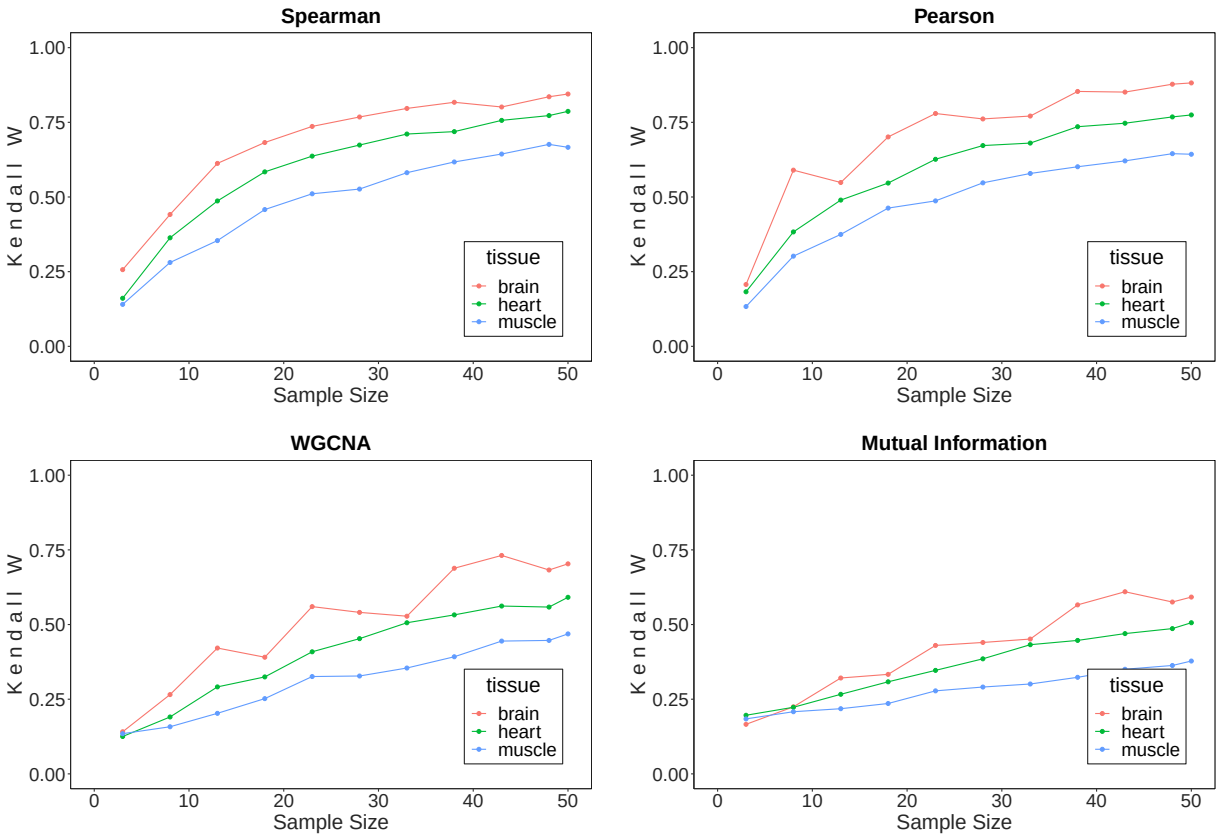


Figure 4.2: Line plots illustrating the results of Kendall concordance coefficient tests for replicate networks constructed from 3 to 50 samples. Each sample size compares 10 replicate data sets constructed from non-overlapping samples of one of three tissues: brain, heart, and skeletal muscle. The plots from left to right show the Kendall W values across sample sizes when constructing networks using Spearman correlation, Pearson correlation, WGCNA, and mutual information, respectively.

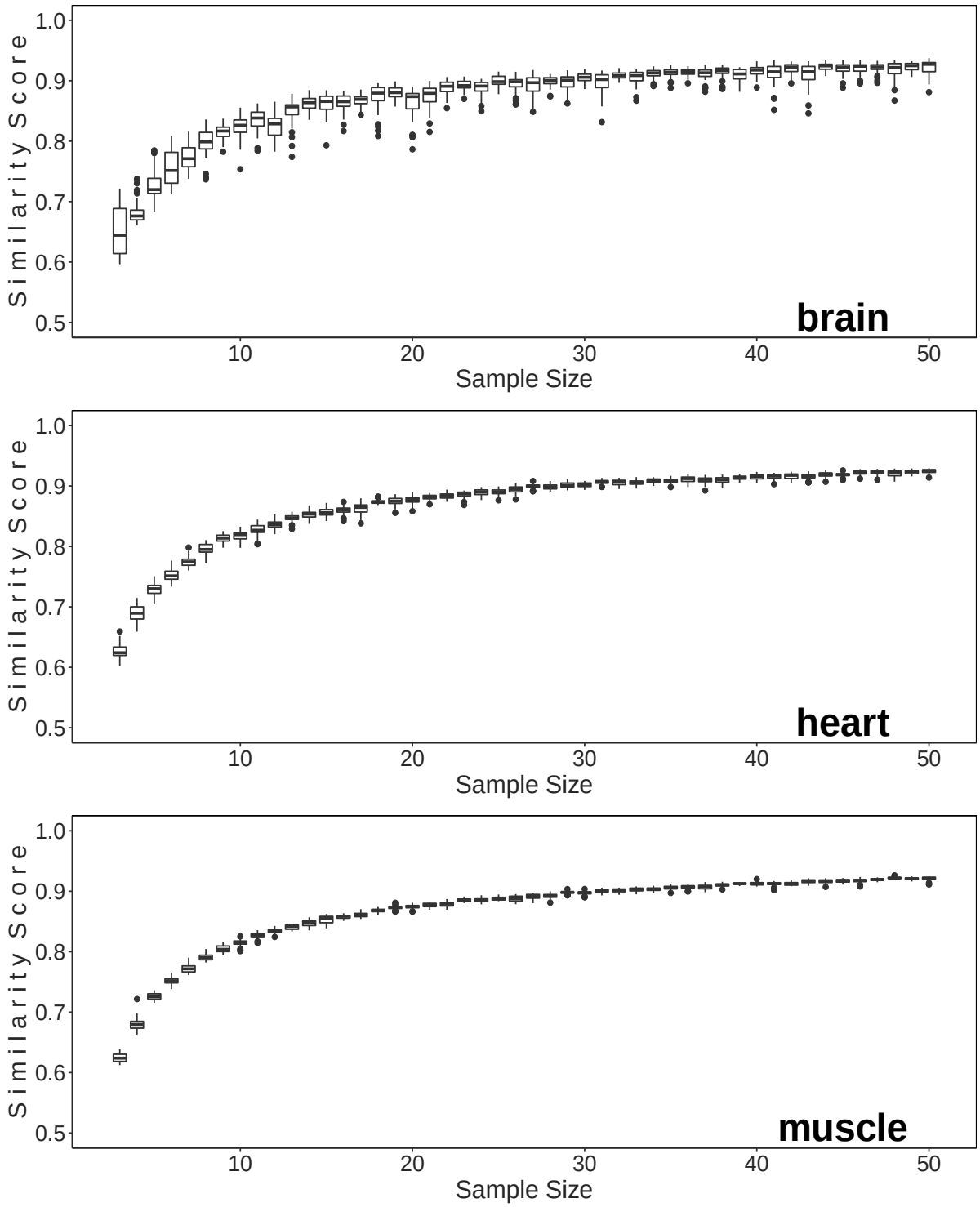


Figure 4.3: Box plots illustrating the results of similarity score calculated using the normalized absolute difference between edge weights between replicate networks constructed using Spearman correlation and from 3 to 50 samples. Each sample size compares 10 replicate data sets constructed from non-overlapping samples of one of three tissues: brain (top), heart (middle), and skeletal muscle (bottom).

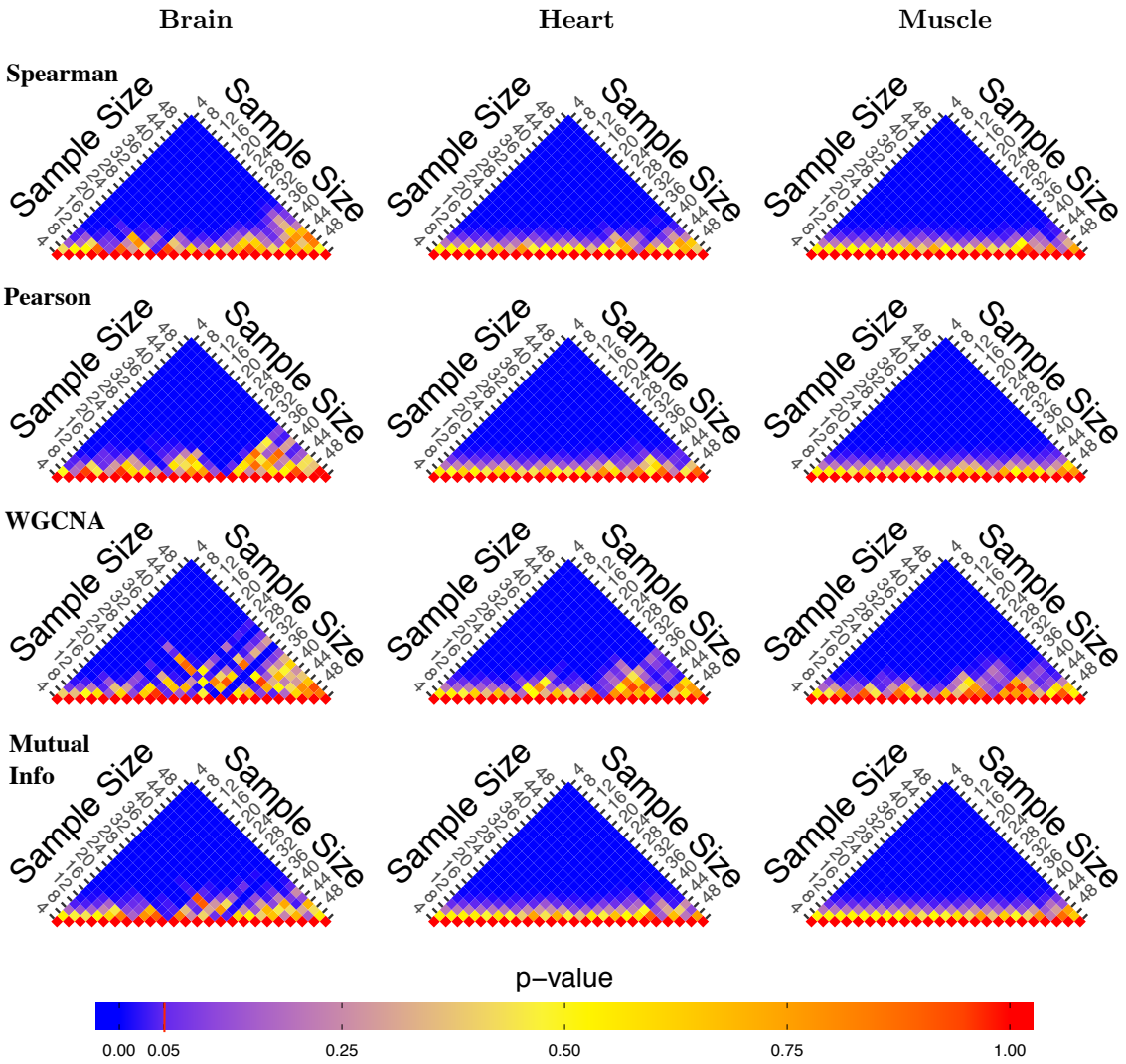


Figure 4.4: Pine forest illustrating the results of the Dunn tests comparing the similarity measure when constructing co-expression networks using sample size $3 \leq s \leq 50$. Only every other sample size result is shown in this plot to summarize the pattern observed as sample size was increased. The left, middle, and right pine plots show the results using samples from brain, heart, and muscle, respectively. In each pine plot, the layers from top to bottom correspond to Pearson correlation, Spearman correlation, WGCNA (signed), and mutual information, respectively. The relationship between the results of the Dunn tests are measured in terms of p-value. The colour blue indicates that the sample sizes being used to construct the co-expression networks result in significantly different edge weights. Any p-value above 0.05, which is marked on the legend at the base of the pine forest, indicates a non-significant difference between the similarity measure of co-expression networks constructed using two different sample sizes. A shade of yellow indicates an intermediate value between 0 and 1, and red indicates a p-value of 1. The red diagonal at the base of each pine plot layer is used as a visual division between layers since the Dunn test was not performed between the results of the same sample size (which would result in a non-significant difference).

We also compared each replicate network to the co-expression networks constructed using all of the available samples for each tissue type across methods. The results were consistent with the results observed when comparing replicate datasets alone. Results using Spearman correlation are presented in Figure B.5 in the Appendix.

4.4 Discussion

In this research, we proposed a generalized method in order to compare the reproducibility of gene co-expression networks. Unlike other strategies that rely on simulated data or functional annotation, this method can be used with real complex gene expression data originating from different phenotypes. We explored how sample size impacts the reproducibility of co-expression networks when constructed with different correlation measures and mutual information as well as using samples originating from different tissues. We quantitatively investigate if the consistency of the co-expression networks is dependent on the method used to construct the networks, the number of samples used to construct the networks, or the sample origin.

From the results of our experiments, utilizing less than 10 samples to construct a network using Spearman correlation, Pearson correlation, WGCNA, or mutual information resulted in relatively inconsistent networks compared to those constructed with a larger number of samples. Another interesting observation was that increasing the number of samples past 10 does not improve the agreement of network similarity score much based on absolute edge weight difference. Indeed, the absolute edge weight difference was high for even lower sample sizes using the GTEx dataset. When assessing the significance of the differences between sample sizes using the Dunn post-hoc analysis, it was observed that the threshold at which the replicate networks are not significantly different is tissue and method dependent. As such, no optimal sample size for all tissues and methods can be established making a methodology that can be used to assess appropriate sample size valuable.

It should be noted that the tissue types converge to a random state of absolute difference similarity (results available upon request) as well. However, no difference is observed in the sample size at which this convergence between tissue types occurs, unlike what was seen in Figure 4.4. Therefore, the increase in reproducibility differences observed between the networks is only a characteristic of the real networks and not due to irrelevant characteristics of the networks.

Furthermore, increasing the sample size beyond 10 samples in the replicate networks does substantially impact the agreement in edge order from the results of Kendall W scores when the edge weights are sorted by weight. This was a relatively consistent pattern across different methods used to construct the networks. However, when utilizing samples from different tissue types, networks constructed from brain samples had Kendall W scores increase more quickly and remain higher than the scores achieved using samples from the other tissue types. This could suggest that reproducibility in the ranked ordering of edges is greatly impacted by the samples used to construct the network. For example, biologically, the gene expression usually varies

less in brain than the other tissues [8]. One consequence of this is that more samples may be required to obtain a reproducible ranked order from samples of one tissue versus another. However, the absolute difference similarity score is also an informative method to capture the difference between two networks (as it measures the difference in weights rather than rankings). All tissues produce quite similar absolute difference similarity scores per sample size. Hence, if one is concerned with the consistency as the sample size increases per tissue, the tissue might not significantly affect results, but additional tissues would be required to say conclusively. However, if one is concerned with orderings of gene correlations, then the tissue is important.

Also, although the similarity score and Kendall W scores achieved were relatively high when constructing networks from 50 samples, it should be noted that neither of the scores utilized ever reach 1, which would mean the same networks were constructed using different groups of samples. One suggestion that could potentially improve the Kendall W scores further would be to bin similar edge weights together. Since correlation could be very similar for groups of edges, if they are slightly shuffled in order, the networks could still be considered highly similar. As such, it would be possible to relax the ranking procedure traditionally used for Kendall rank coefficient to allow similar edge weights to be considered tied in rank. To do so, one could select a threshold that establishes how close the co-expression values can be across edges to be considered tied in rank. With real biological data, it appears unlikely that entirely reproducible networks are likely to be constructed. However, the scores utilized obtain high enough scores that any improvement obtained by adding more samples is likely marginal.

One limitation of this study is that non-overlapping datasets were not generated past 50 samples for all three tissues used. However, analysing non-overlapping or slightly overlapping datasets with sample sizes larger than 50 is suggested as future work for different datasets including phenotypes that have high heterogeneity. Furthermore, the method proposed in this paper should be applicable to any gene expression that a gene co-expression network can be constructed. As future work, determining if the organism type also impacts the number of samples required similar to tissue type would be of interest for those performing cross-species gene co-expression network studies.

The common assumption that 20 samples are enough for generating reproducible co-expression networks has commonly been justified based on well-behaved normally distributed data. However, our observation suggests that in the context of real-world data with various degrees of heterogeneity, there is no optimal one-size-fits-all solution. Therefore, we suggest utilizing the proposed methodology to estimate the optimal sample size for a given method and phenotype of interest. Leveraging publicly available datasets, a researcher can find datasets of comparable heterogeneity (or phenotype) to estimate the sample size required for an experiment.

4.5 Conclusion

In this research, we utilize a systematic, quantitative approach to study the effect of sample size on the reproducibility of co-expression networks construction using several correlation measures and RNA-seq data from different tissue types. Two measures were utilized to determine if the consistency of networks is more dependent on method used to construct the networks, sample origin, or the number of samples used to construct the networks. Our results showed that the consistency of co-expression networks increases as sample size increase, but is relatively high by approximately 10 samples; any additional samples only considerably improves the ranked order of the edges rather than the graphs overall. The difference between ranked orders differ by tissue type, which is the case for all methods used to construct the networks. This means that if the ranked order of gene correlations is important, the tissue type used to construct the networks could be important.

References

[1] Fadhl M Alakwaa. Modeling of gene regulatory networks: a literature review. *Journal of Computaional Systems Biology*, 1:1–8, 2014.

[2] Jeffrey D Allen, Yang Xie, Min Chen, Luc Girard, and Guanghua Xiao. Comparing statistical methods for constructing large scale gene networks. *PloS One*, 7(1):e29348, 2012.

[3] G Altay. Empirically determining the sample size for large-scale gene network inference algorithms. *IET Systems Biology*, 6(2):35–43, 2012.

[4] Gerald J Bakus. *Quantitative Analysis of Marine Biological Communities: Field Biology and Environment*. John Wiley & Sons, 2007.

[5] Sara Ballouz, Wim Verleyen, and Jesse Gillis. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*, 31(13):2123–2130, 2015.

[6] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[7] Latarsha J Carithers, Kristin Ardlie, Mary Barcus, Philip A Branton, Angela Britton, Stephen A Buia, Carolyn C Compton, David S DeLuca, Joanne Peter-Demchok, Ellen T Gelfand, et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreservation and Biobanking*, 13(5):311–319, 2015.

[8] Esther T Chan, Gerald T Quon, Gordon Chua, Tomas Babak, Miles Trochesset, Ralph A Zirngibl, Jane Aubin, Michael JH Ratcliffe, Andrew Wilde, Michael Brudno, et al. Conservation of core gene expression in vertebrate tissues. *Journal of Biology*, 8(3):33, 2009.

[9] Djordje Djordjevic, Andrian Yang, Armella Zadoorian, Kevin Rungrugeecharoen, and Joshua WK Ho. How difficult is inference of mammalian causal gene regulatory networks? *PLoS One*, 9(11), 2014.

[10] Shang Gao, Abdullah Sarhan, Reda Alhajj, Jon Rokne, Doug Demetrick, and Jia Zeng. Quantifying gene co-expression heterogeneity in cancer towards efficient network biomarker design. *Current Bioinformatics*, 10(3):306–314, 2015.

[11] Florian Geier, Jens Timmer, and Christian Fleck. Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC Systems Biology*, 1(1):11, 2007.

[12] Zachary F Gerring, Eric R Gamazon, Eske M Derks, et al. A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression. *PLoS Genetics*, 15(7):e1008245, 2019.

[13] Alexander J Hartemink. Reverse engineering gene regulatory networks. *Nature Biotechnology*, 23(5):554–555, 2005.

[14] OD Iancu, P Darakjian, B Malmanger, NAR Walter, Shannon McWeeney, and Robert Hitzemann. Gene networks and haloperidol-induced catalepsy. *Genes, Brain and Behavior*, 11(1):29–37, 2012.

[15] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, 2008.

[16] Franziska Liesecke, Johan-Owen De Craene, Sébastien Besseau, Vincent Courdavault, Marc Clastre, Valentin Vergès, Nicolas Papon, Nathalie Giglioli-Guivarch, Gaëlle Glévarec, Olivier Pichon, et al. Improved gene co-expression network quality through expression dataset down-sampling and network aggregation. *Scientific Reports*, 9(1):1–16, 2019.

[17] Wei Liu, Liping Lin, Zhiyuan Zhang, Siqi Liu, Kuan Gao, Yanbin Lv, Huan Tao, and Huaqin He. Gene co-expression network analysis identifies trait-related modules in Arabidopsis thaliana. *Planta*, 249(5):1487–1501, 2019.

[18] Farhad Maleki, Katie Ovens, Daniel J Hogan, and Anthony J Kusalik. Gene set analysis: Challenges, opportunities, and future research. *Frontiers in Genetics*, page (To Appear), 2020.

[19] Farhad Maleki, Katie Ovens, Ian McQuillan, and Anthony J Kusalik. Size matters: how sample size affects the reproducibility and specificity of gene set analysis. *Human Genomics*, 13(1):42, 2019.

[20] Farhad Maleki, Katie Ovens, Ian McQuillan, Elham Rezaei, Alan M Rosenberg, and Anthony J Kusalik. Gene set databases: A fountain of knowledge or a siren call? In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 269–278. ACM, 2019.

[21] Michael Neidlin, Smaragda Dimitrakopoulou, and Leonidas G Alexopoulos. Multi-tissue network analysis for drug prioritization in knee osteoarthritis. *Scientific Reports*, 9(1):1–12, 2019.

[22] Katie Ovens. Measuring Consistency of Co-expression Networks. http://github.com/klovens/compare, 2019.

[23] Katie L Ovens, Daniel J Hogan, Farhad Maleki, Ian McQuillan, and Anthony J Kusalik. pineplot: an R package for visualizing symmetric relationships. In *Proceedings of the Tenth International Conference on Computational Systems-Biology and Bioinformatics*, pages 1–8, 2019.

[24] Ralf Steuer, Jürgen Kurths, Carsten O Daub, Janko Weise, and Joachim Selbig. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl_2):S231–S240, 2002.

[25] Paolo Tieri, Lorenzo Farina, Manuela Petti, Laura Astolfi, Paola Paci, and Filippo Castiglione. *Network Inference and Reconstruction in Bioinformatics*. Elsevier, 2019.

[26] Sipko van Dam, Urmo Vosa, Adriaan van der Graaf, Lude Franke, and Joao Pedro de Magalhaes. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, 19(4):575–592, 2017.

[27] Veronica Vinciotti, Ernst C Wit, Rick Jansen, Eco JCN de Geus, Brenda WJH Penninx, Dorret I Boomsma, and Peter AC't Hoen. Consistency of biological networks inferred from microarray and sequencing data. *BMC Bioinformatics*, 17(1):254, 2016.

Chapter 5

Juxtapose: A Python tool for gene embedding for co-expression network comparison

In this chapter, the genes of different gene co-expression networks were mapped to vectors in order to represent the genes of each network in a lower-dimensional space; i.e., create a gene embedding based on gene expression data. By doing so, distances between genes of interest could be calculated and used as a representation of how similar or dissimilar the genes are when comparing genes within or between gene co-expression networks. Not only this, we created a joint embedding so that multiple co-expression networks could be compared. This allows for the calculation of global and local similarities between the networks to be calculated more efficiently than a method such as network alignment.

The development of Juxtapose addresses many of the concerns brought up in Chapter 2, including the difficulties with comparing more than one-to-one orthologs, considering GCN edge weights, handling large or dense GCNs, and comparing a large number of species for evolutionary studies. The capabilities of Juxtapose are further demonstrated in Chapter 6.

This paper has been submitted to the journal *BMC Bioinformatics*.

Citation

Katie Ovens, Farhad Maleki, B Frank Eames, and Ian McQuillan. 2020. Juxtapose: A Python tool for gene embedding for co-expression network comparison. Submitted to *BMC Bioinformatics*.

Author contributions

Katie Ovens wrote the paper and wrote the code for the initial GCN walk generation and embedding software. Katie Ovens was also responsible for performing all evaluations in the paper—e.g. generating synthetic networks, constructing networks from real data, making network comparisons, writing the code to make visualizations, biclustering analyses, and enrichment analyses, etc. Farhad Maleki helped with optimizing, parallelizing, and refactoring the code for GCN walk generation and gene embedding. He also helped to revise the manuscript. Brian Eames and Ian McQuillan co-supervised the work and assisted with revision of the manuscript.

Abstract

Gene co-expression networks (GCNs) are not easily comparable due to their complex structure. In this paper, we propose a tool, Juxtapose, together with similarity measures that can be utilized for comparative transcriptomics between a set of organisms. While we focus on its application to comparing co-expression networks across species in evolutionary studies, Juxtapose is also generalizable to co-expression network comparisons across tissues or conditions within the same species. A word embedding strategy commonly used in natural language processing was utilized in order to generate gene embeddings based on walks made throughout the GCNs. Juxtapose was evaluated based on its ability to embed the nodes of synthetic structures in the networks consistently while also generating biologically informative results. Evaluation of the techniques proposed in this research utilized RNA-seq datasets from GTEx, a multi-species experiment of prefrontal cortex samples from the Gene Expression Omnibus (GEO), as well as synthesized datasets. Biological evaluation was performed using gene set enrichment analysis and known gene relationships in literature. We show that Juxtapose is capable of globally aligning synthesized networks as well as identifying areas that are conserved in real gene co-expression networks without reliance on external biological information. Furthermore, output from a matching algorithm that uses cosine distance between GCN embeddings is shown to be an informative measure of similarity that reflects the amount of topological similarity between networks. A development version of the software used in this paper is available at *https://github.com/klovens/juxtapose*.

5.1 Introduction

High-throughput techniques such as RNA-seq and microarray make it possible to measure the expression level of a large number of genes in a single experiment. These high-throughput expression studies have resulted in a large number of gene expression datasets that are available through public repositories such as GEO [11] and ArrayExpress [4]. Differential expression analysis, which refers to the comparison of the expression measures of individual genes across phenotypes/conditions, has been the common practice in analysing these data [2]. This approach only leads to the identification of individual genes with different expression levels across phenotypes/conditions. However, often coordinated interaction of groups of genes drives various biological processes and functions, and the change in the expression level of a single gene does not capture this complex network of interactions. These complex gene-gene interactions can be modeled as a network.

Networks have been widely used for the study of complex interactions between genes, proteins, and other biomolecules [37, 43]. In particular, gene co-expression networks (GCN) constructed using gene expression data can be utilized to extract information about coordinately expressed genes. It has been shown that co-expression networks are not static, and can change depending on the the biological context [35]. Comparing these networks can aid in improving functional annotation of genes and the discovery of gene–gene interac-

tions [13], revealing the molecular mechanisms of complex diseases or the relationships between biological processes [21], and helping to speed up the process of selecting genes for targeted mutational studies [35]. Therefore, comparing these networks can provide valuable insight into the key coordinated interactions that are associated with the phenotypes under study.

Weighted gene co-expression network analysis (WGCNA) [21] is one of the methods most commonly used to study the relationships between co-expression modules and to test whether a module is preserved between two different phenotypes. Although WGCNA provides insight into the conserved modules between the pairwise comparison of phenotypes, it does not provide a systematic means for comparing more than two phenotypes or networks. OrthoCluster [42] is another method that can be used to align modules in a pairwise comparison of phenotypes. However, it relies on external biological information such as one-to-one orthologs that is not always readily available specifically for non-model organisms [30]. Furthermore, different genes throughout evolution can take on similar roles and processes [1, 28, 41], and matching orthologs is not always appropriate when comparing GCNs. In contrast, it is possible to compare networks by strictly using the topology of the networks. However, comparing co-expression networks topologically is challenging due to their large size and the computational complexity of this type of network comparison [17]; therefore, the application of network comparison strategies such as network alignment—more commonly applied to protein-protein interaction (PPI) networks—to larger GCNs might be difficult.

Embedding techniques, a powerful tool in natural language processing, have also been utilized to analyse biological networks. These include matrix factorization-based methods as well as more recent neural network-based methods [12, 16]. Embedding methods provide a vectorized representation for each gene/protein and are often faster than other options, which can be critical when dealing with analysing networks [38]. Additionally, the learned embeddings are often applicable for downstream analysis as the method provides a numeric representation of the genes that can be fed into a machine learning algorithm, for example, while capturing information about how it is positioned in a network.

In this paper, we present Juxtapose, a systematic methodology for comparing multiple co-expression networks using an embedding-based approach. The proposed method does not require external biological information such as knowledge of orthologs. Juxtapose establishes both a local and global measure of similarity between networks based on their topology. Using both synthesized and real networks, we show the utility of the proposed method for comparing GCNs. In the lack of network alignment methods specialized for GCN alignment, we compare to PPI alignment methods that have been used or can be used to compare GCNs [13, 25]. We also compare Juxtapose to MUNK, which has many similarities with our proposed embedding method. However, it has been designed for PPI alignment, so it is unknown how well it performs when aligning GCNs. Furthermore, the biological relevance of the gene set enrichment analysis results after aligning real GCNs from multiple species using Juxtapose is compared to the results obtained using a common method used to compare GCNs, WGCNA.

The rest of the paper is structured as follows. Related work on GCN and PPI network analysis using

embedding is described in Section 5.2. Section 6.2 presents the methodology in detail and Section 6.3 describes the results obtained when comparing GCNs. Section 5.5 discusses the results and identifies potential caveats. Finally, Section 5.6 ends the paper with a brief summary.

5.2 Related Work

Embedding methods stem from natural language processing (NLP), a discipline concerned with the computational methods for understanding and analysing text. An embedding for a word is a vectorized representation, i.e. a point in embedding space. Methods for learning embeddings rely on the Distributional Hypothesis, which states that words that appear in the same contexts share semantic meaning [18]. As such, semantically similar words should be mapped close to each other in the embedding space. In terms of embedding genes in the context of GCNs, co-expressed genes should be placed close together in the embedding space.

Word2vec is a neural network-based approach, which aims at learning a distributional representation of words as vectors [26]. The key components of this model are two weight matrices. The rows of the first matrix and the columns of the second matrix embed the input genes and target genes, respectively. The product of these two gene vectors is then used to get the probabilities for being a target gene, given the selected input word. A gradient descent approach can be used to learn these weight matrices by maximizing the probabilities of the true target gene(s).

Methods that extend or utilize word2vec to embed graphs such as node2vec [16] generate random walks through the networks to generate node representations. When embedding GCNs, a sequence of genes can be generated by conducting a random walk on the network. These walks capture the organization of the genes in the GCN e.g., the more two genes appear in sequence, the closer their gene embedding representations will become during the model training process. However, as node2vec was not designed to consider networks with edge weights and also does not offer strategies to create embeddings to compare across networks, we did not make use of the pipeline directly for graph embedding as it would ignore essential characteristics of GCNs.

Recent advances in machine learning have led to the development of gene representations from co-expression networks [6, 7, 10]. Gene2vec [10] and G2vec [6] are examples that utilize the word2vec [26] model originally used for natural language processing. Word2vec aims to predict the co-occurrence of a word and its surrounding words, which is called the context for that word. Analogously, in GCNs genes that are co-expressed with a given gene are considered its context. Knowing a gene and its context, these methods try to predict a gene from its context or vice versa.

Currently, these techniques have been used to predict important genes for disease within a single co-expression network. Gene2vec [10], utilizes word2vec as well as a measure of "clusteredness" of known biological pathways from MSigDB to learn gene embeddings. They used the "clusteredness" measure to encourage genes that are part of the same biological process or function to cluster together in the embedding

space. They evaluated their method by its capability to cluster genes in the same biological categories, as defined by MSigDB. G2vec [6] also used word2vec to compute gene representations for identifying potential biomarkers important for cancer prognosis. Using gene expression data from cancer patients, the authors divided samples into two groups of poor and good prognosis as defined by survival outcome. For each group, they built a GCN. Then for each GCN, they generated random walks (10 walks originating from each gene). Next, these random walks were used for learning gene representations that distinguish good and poor prognosis groups. Using gene expression data acquired from TCGA transcriptomic dataset, Choy et al. implemented a two-layer neural network architecture to learn gene representations from cancer biomarker discovery [7]. To learn an association between the category of each sample and its gene expression, they trained the model to minimize the error between the predicted and actual gene expression values. They evaluated their model by its capability in clustering similar samples in the embedding space. G2vec [6] is the only method of those described above that directly compared two networks in a pairwise manner. However, combining walks from different GCNs to train a single model will convolute the gene representations as they will be a mixture of both networks. Furthermore, all of these methods utilize random walks as is traditionally done when embedding networks, which does not incorporate the weights of the edges in GCNs.

Fan et al. used a matrix factorization method as well as one-to-one orthologs to compare PPI networks of well-studied species, namely human, mouse, and two types of yeast [12]. Given a source PPI network, a target PPI network, and a set of homologous proteins across species, they computed diffusion kernels for each PPI network. Next, the diffusion kernel for the source species is factorized. To create protein representations that embed proteins from different species to the same embedding space, they solved a linear system of the source and target species' diffusion kernels. The choice of the homologous proteins is essential for this approach as it can substantially affect the results of the linear system used for enforcing the embedding of multiple species to the same embedding space as it is a hard condition when solving the linear system [27].

5.3 Methods

In this section, the following will be described: the synthetic and real datasets used, the GCN construction and methodology of Juxtapose, the evaluation, and a comparison to other approaches in the literature.

5.3.1 Data

To analyse the accuracy of the results of the proposed method, Juxtapose, we use synthetic and real GCNs. The 3 synthetic networks are shown in Figure 5.1, which are only evaluated to test each method's ability to align identical networks. For the real datasets, we utilized RNA-seq data available from the GTEx project, which has expression data across many different tissues. To construct GCNs for brain and heart tissues, we used subsets of the expression data from heart ($n = 200$ samples) and brain tissue ($n = 200$ samples). Gene expression and sample description data were downloaded on January 18th, 2020 from the GTEx website. We

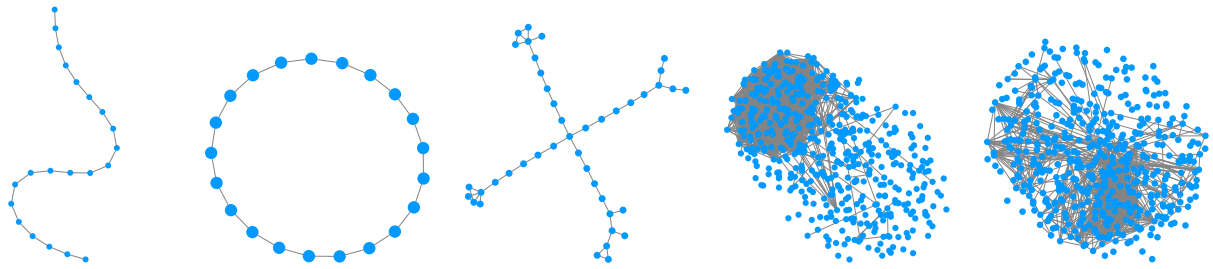


Figure 5.1: Networks used for evaluating Juxtapose. The line, circle, and cross were synthetic networks and the last two networks are a heart and brain GCN, respectively.

Table 5.1: Gene sets used for constructing co-expression networks

ID	Description	# of Genes
hsa04260	Cardiac muscle contraction	87
hsa05410	Hypertrophic cardiomyopathy	90
hsa05010	Alzheimer disease	369
hsa05012	Parkinson disease	249
Potential Anchor Genes		
GO:0019725	cellular homeostasis*	970

*GO gene set was used to select candidate anchor genes

used a common pipeline for preprocessing RNA-seq data [22]. The preprocessing was conducted by using Trimmed Mean of M-values (TMM) normalization, and filtering lowly expressed genes was done using the *edgeR* [34] and *limma* [33] packages in R. Several KEGG pathways in humans—see Table 5.1 for the list of pathways—were selected in order to construct the networks from brain and heart tissues using a method discussed in Section 5.3.2. It is hypothesized that the GCNs constructed from heart tissue samples would have more conserved networks when these GCNs are compared to each other than when compared to GCNs constructed using brain tissue samples. Similarly, brain GCNs should show more similarities to each other. Two of the networks constructed are shown in Figure 5.1. Lastly, we also utilized an RNA-seq dataset originating from the prefrontal cortex of human, chimpanzee, macaque, and mouse [3] to evaluate Juxtapose using a multi-species dataset. This dataset contained 12 samples for each species. The reads of this dataset were mapped to Ensembl genome builds GRCh38, Pan_tro_3.0, Mmul_10, and GRCm38 using STAR 3.5.2 [9]. The raw counts were normalized for each species individually using TMM normalization. Any gene that did not meet thresholds was removed from downstream analyses.

To construct the real GCNs, Pearson Correlation Coefficient (PCC) was calculated for each pair of genes and transformed to a value between 0 and 1 using $0.5 + 0.5PCC(g_i, g_j)$ where g_i and g_j are a pair of genes in a network G. Although PCC ranges from -1 (negative correlation) to 0 (no correlation) to $+1$ (positive correlation), the affine transformation above was applied to map negative correlation to 0, no correlation to 0.5, and positive correlation to $+1$. This ensures that negative correlations are separate and preserved, while

allowing the values to be between 0 and 1. In order to construct the co-expression networks, a threshold of $+/-0.8$ for the original PCC values was selected before being transformed to determine whether an edge/relationship should connect a pair of genes.

5.3.2 Projecting genes from different networks into the same embedding space

When networks are embedded separately, they are not necessarily going to be directly comparable. Therefore, it is useful to have pieces of the networks with a known and conserved structure. In this way, these pieces can be matched up with high confidence and can be used to align the parts of the network with unknown topology.

One strategy that has been used to jointly embed multiple PPI networks is to use a group of landmark or anchor genes [12]. However, there is an important difference when doing this procedure with co-expression networks. The selection of anchor genes when comparing co-expression networks is critical since expression relationships between some orthologous genes can vary widely depending upon the phenotypes or organisms being compared. To avoid this problem, anchor genes were selected from highly-conserved cellular processes, such as transcription and translation, which more likely contain orthologous gene positions within the co-expression network [40]. Therefore, anchor genes were selected from those annotated with homeostatic processes involved in the maintenance of an internal steady state at the level of the cell, including control of cellular proliferation and death and control of metabolic function. These genes were selected from gene sets shown in Table 5.1. These genes are likely to have similar connections to the rest of a co-expression network. In order to compare co-expression networks using these anchor genes, we propose a method to embed genes in the same embedding space.

Figure 5.2 illustrates the steps for preparing two co-expression networks G_1 and G_2 for embedding and these steps are described as follows. First, the anchor genes need to be selected, $a_1, ..., a_n$, that are present in the networks that will be embedded. We use anchor genes to provide a base for model evaluation. Since anchor genes are expected to be aligned across species, if we have the same graph structure attached to these genes, the structures also are expected to be aligned across GCNs. Thus, different synthetic structures are created and the same structure is added to matching anchor genes across networks. Hereafter, we refer to such a synthetic structure as a dangling structure. For a selected anchor gene, the dangling structure created is a random sparsely connected graph. The number of nodes for a dangling structure γ ($\gamma \in \mathbb{N}$) is a hyperparameter for the model representing the number of nodes in a dangling structure. The nodes in the dangling structure are connected using 15% of all its potential edges. If the resulting dangling structure is not a connected graph, a minimal number of edges required for making the dangling structure a single component is randomly added to the dangling structure. All edges in the dangling structure are assigned a weight equal to 1.

The rationale behind using a sparse artificial network is that nodes in dense networks are topologically similar. Therefore there is not much topological variation that can be used for model evaluation, and any

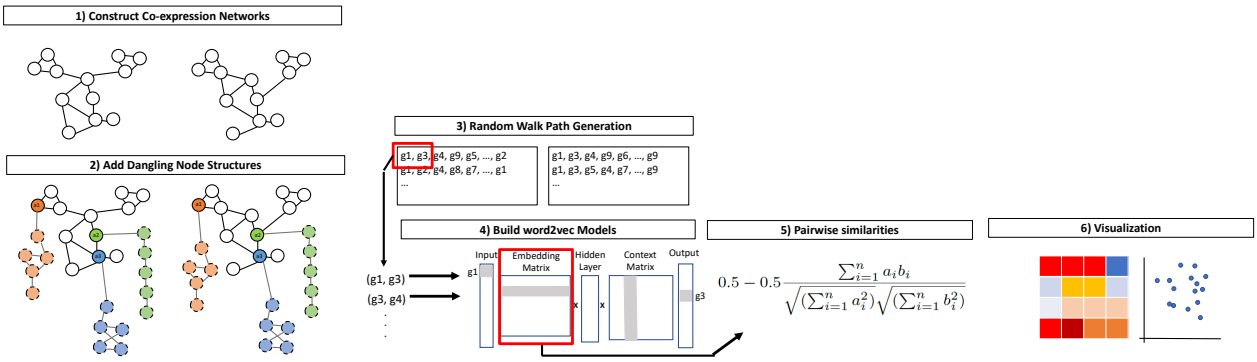


Figure 5.2: Methodology for generating joint gene embeddings from co-expression networks. (1) The co-expression networks are constructed from gene expression data. (2) Anchor genes $(a_1, ..., a_n)$ are selected as anchor nodes, which have relatively stable behaviour in the co-expression networks being compared. Dangling structures of γ artificial nodes are added to the graphs (shown with dashed borders and edges shown in grey) with equal edge weights across the networks being compared. In this illustration $\gamma = 4$. These dangling structures are connected to one of the selected anchor nodes in the original networks. (3) These networks are used to generate a set of random walks from each gene in each network. (4) The paths through the nodes are used as sentences to feed to a word2vec model, which learns informative embeddings for each gene in the networks. The model takes a gene in a network and the genes surrounding it in a path within a defined window and feeds them to a neural network that, after training, predicts the probability that each gene appears in the window around the focus gene. The process begins with a vector that contains all zeros and a 1 which represents the corresponding gene in the network. An $N \times \|G\|$ embedding matrix contains one row for every gene in the vocabulary and the number of columns equal to the embedding size N. Pairs of genes are used to train the model and generate a representative embedding for each gene. This newly discovered embedding vector of dimension N forms the hidden layer. The input gene, selected using multiplication of the embedding matrix and Input vector, is fed to the model. The multiplication of the hidden layer and the word context matrix produces the output, which will be a prediction of the most probable output gene. Then, the loss is calculated between what was expected and the gene predicted. During backpropogation, when computing the gradient of the loss function, network weights including the embeddings for all genes in the vocabulary get updated. Given a hypothetical path from a random walk $g1, g3, g4, g9, \ldots, g2$ and a window size of 2, $g3$ has the following input gene pairs $(g1, g3)$ and $(g3, g4)$ under the Skip-gram architecture of word2vec. (5) The pairwise similarity scores between genes in the embedding matrix are calculated resulting from the word2vec model. (6) The embeddings and the distances between genes in the embedding are are analysed and visualized.

naive embedding that maps all nodes to almost the same value could be considered a reasonable solution. Therefore, we use sparse graphs to provide a better estimate of model capability in encoding topological variation among nodes.

5.3.3 Generating walks and model training in Juxtapose

Walk generation was performed in Juxtapose by converting the weights of the GCNs to a probability of travelling through the edges connecting genes. The higher a correlation value, the more likely a walk would travel through the edge. In order to handle large real networks in Juxtapose, translation from gene names or Ensembl IDs to integer values was performed in order to give the method the power to generate a large number of walks quickly. This translation generates a JSON file to make it convenient to convert integer

values back to gene name or IDs for visualization and for interpreting results.

In order to generate gene embeddings, *gensim* version 3.8.3 was utilized. A word embedding was trained by maximizing the probability of gene co-occurrences in context, i.e., only a few genes apart in a single walk. Analogously, we defined the context of a gene by the other genes that are co-expressed with it. An $N \times \|G\|$ embedding matrix is randomly initialized and contains one row for every gene in the vocabulary and the number of columns equal to the embedding size N. This newly discovered embedding vector of dimension N forms the hidden layer. Pairs of input genes and context genes, i.e., expected output, are generated using a sliding window through each walk that was made for the genes in the GCN. An input gene is fed to the model in order to generate a prediction of possible context genes. The multiplication of the hidden layer and the word context matrix produces a prediction of the most probable context gene(s). Then, the loss is calculated between what was expected and the gene(s) that were predicted. This process continues with all of the generated walks and subsequent training samples with the model being iteratively updated to make better predictions for context genes based on an input gene.

The parameters used to generate the embeddings for genes in each dataset are provided in Table C.1 of the Appendix. We rely on our ability to generate training data from the GCNs by using more walks per gene rather than increasing the number of training epochs or iterations, which can cause overfitting [15]. This is often not possible for many applications since the amount of training data can be limited. However, in the context of GCN, one can extract a large dataset of random walks. This has the benefit of (1) providing a better representation of a co-expression network by having a large number of random paths and (2) not needing to repeat the training for a large number of epochs. Indeed, our model used only 1 epoch, and it generated a large number of random paths from the entire network. Instead of, for example, using 10 walks per gene and iterating over this 100 epochs we use 1000 walks per node with 1 epoch.

5.3.4 Measuring similarity of embedded genes, aligning networks, and measuring network similarity with Juxtapose

One local and one global measure of similarity between genes was used in order to compare the genes of two co-expression networks. The local similarity measure utilized between all pairs of gene vectors between the two networks was cosine distance, which measures the cosine of the angle between them. Cosine distance between pairs of genes was calculated as shown in Equation 5.1, where $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$ are the gene vectors/embeddings.

$$\text{cosine distance}(a, b) = 0.5 - 0.5 \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{(\sum_{i=1}^{n} a_i^2)}\sqrt{(\sum_{i=1}^{n} b_i^2)}} \tag{5.1}$$

One advantage of cosine distance is that it has low computational complexity, where only the non-zero dimensions of the gene vectors need to be considered. Furthermore, cosine distance tends to be effective at estimating the distance between vectors when they have a high dimension [14]. Indeed, as the structure of GCNs can be quite complex, and the number of genes in these networks is often in the thousands, the gene

embeddings may require a high dimension in order to represent their position in the GCNs accurately.

With this local distance measure between genes of different networks, it is then possible to match genes from one network to the other. A matching algorithm (formally, on bipartite graphs) is an algorithm that takes two lists of elements where there is a distance between every element of one list to every element of the other, and constructs a "matching" between the two lists—a matching associates every element of one list with exactly one element of the other list in such a way that each element only gets associated once—and it does so in such a way that the sum of the distances matched is minimal over all possible associations. The Hungarian algorithm is a well-known matching algorithm that runs in polynomial time complexity. The matching constructed by the algorithm is mathematically guaranteed to be optimal, and have the smallest sum of matched distances [20]. In our case, the two lists are the genes in the two GCNs being compared, and the distance between pairs of genes of the two networks being compared is the cosine distance. Thus, the Hungarian algorithm in the scikit-learn Python library [31] is used to create a type of global similarity by producing the best global alignment (matching) of genes in two networks based on their pairwise angular distance. This matching not only provides an optimal association (or alignment) between genes of the two networks, but the sum (or equivalently, average) of the matched distances provides a global similarity score between the networks being compared. As there was a distance calculated between each pair of genes, groups of genes that have similar patterns of distances can also be grouped using a biclustering method. This can also be overlaid with other biological information for other downstream analyses.

Biclustering was utilized in order to discover groups of genes that have similar distances to each other as well as similar differences to other genes. Spectral Biclustering assumes a checkerboard structure where the same gene can belong to multiple biclusters [19]. The rows and columns of a matrix with this structure may be partitioned so that the entries of any bicluster in the Cartesian product of row clusters and column clusters are approximately constant. For instance, if there are two row partitions and three column partitions, each row will belong to three biclusters, and each column will belong to two biclusters. Biologically, genes may be involved in different biological processes and have different patterns of distance between genes. This method of biclustering was used since the biclusters generated provide clusters of genes that have similar distances from a gene of interest to different degrees. Gene set analysis was performed on the resulting biclusters on the non-simulated networks using WebGestalt [44].

5.3.5 Evaluation of Juxtapose

Results from two common methods of graph alignment, IsoRankN [23] and MAGNA++ [39], as well as MUNK [12] and WGCNA [21] were compared to the results of the gene embedding method Juxtapose, where appropriate. IsoRankN and MAGNA++ were evaluated based on their ability to align the nodes of equal or similar networks and the information captured by their similarity scores. Real networks for brain and heart were also compared to each other in order to compare similarity results from Juxtapose to the results from IsoRankN, MAGNA++, and MUNK. The percentage of correctly aligned genes was determined by

measuring the proportion of genes with corresponding gene names in each aligned GCN that were matched together in an alignment. Juxtapose was further evaluated with large, real GCNs from multiple organisms to demonstrate its ability to handle various GCNs with different genes as well as to assess the method from a biological perspective. WGCNA was compared based on the conserved modules identified in pairwise comparisons between GCNs of real large networks from the prefrontal cortex of multiple species.

5.4 Results

The following sections present the results of network comparison using Juxtapose. Section 5.4.1 reports the results of comparing identical synthetic and real GCNs using Juxtapose and comparing these results to PPI network alignment methods IsoRankN and MAGNA++. Section 5.4.2 includes the comparison of GCNs constructed using different subsets of samples from brain and heart tissue samples and compares the results of Juxtapose to IsoRankN, MAGNA++, and MUNK. Finally, Section 5.4.3 applies Juxtapose to large GCNs constructed from multiple species and compares to WGCNA.

5.4.1 Alignments of identical networks

Table 5.2 indicates the percentage of correctly matched genes for IsoRankN, MAGNA++, and Juxtapose. Both IsoRankN and MAGNA++ have a parameter (alpha) that for IsoRankN indicates the extent to which network topology is used to make the network alignment—where 1 is completely topology based—and MAGNA++ has a alpha value that balances between node and edge conservation. Furthermore, we provide these methods with different degrees of knowledge about known node matches between the networks in the form of (sequence similarity) bitscores. If 100% of the bitscores are provided, this means that the bitscores clearly indicate which matches are the most appropriate matches between nodes e.g. the corresponding genes between networks have a value set to 1 and the remaining node matches are set to zero. Juxtapose does not used any sequences and therefore matching does not take sequence similarity into account and is purely topologically based. The performance of the alignment methods was measured based on their ability to align the corresponding genes in the networks compared e.g., $gene_1$ in a GCN correctly aligned to $gene_1$ in a duplicate version of the GCN would be counted as a match.

IsoRankN and MAGNA++ were able to match all of the corresponding nodes of the two networks only when provided with the known matches in the form of high sequence similarity i.e., high bitscore values. This is in agreement with an observation also reported by Singh et al. [36] that including sequence information improves the performance significantly. These methods sometimes struggled with aligning the structures that had symmetry such as the line and circle synthetic networks if artificial bitscore matches were not provided to the algorithms. IsoRankN had relatively higher scores than MAGNA++ for the synthetic networks when no biological similarity was used during the alignment process. The exception of low performance without bitscores was that MAGNA++ was able to align the heart GCN with 93% of the nodes matched correctly.

Table 5.2: Percentage of matched genes in self-aligned networks reported for MAGNA++, IsoRankN, and Juxtapose. The alpha values indicate the balance between node similarity and edge similarity (MAGNA++) or the balance between topological similarity and sequence similarity (IsoRankN). When no percentage of bitscores is provided, the algorithm was not provided with informative bitscores i.e. the match between any genes was equally likely. When 50% of bitscores were provided, 50% of the genes had the highest bitscore provided for the real match between the genes of both networks. When 100% of bitscores were provided, 100% of the genes had the highest bitscore provided for the real match between the genes of both networks. N/A is given for the settings in Juxtapose as no bitscore file is provided and no alpha value is provided to the tool.

	MAGNA++				**IsoRankN**				**Juxtapose**
	alpha 0.50	alpha 0.95	alpha 0.50 with 50% bitscores	alpha 0.50 with 100% bitscores	alpha 0.50	alpha 0.95	alpha 0.50 with 50% bitscores	alpha 0.50 with 100% bitscores	**N/A**
Line	0	0.10	0.52	1.0	0.24	0.19	0.0	1.0	1.0
Circle	0	0.14	0.57	1.0	0.33	0.29	0.0	1.0	1.0
Cross	0.24	0.02	0.52	1.0	0.29	0.19	0.83	1.0	1.0
Heart	0.93	0.93	0.99	1.0	0.16	0.04	0.71	1.0	1.0
Brain	0.33	0.55	0.99	1.0	0.18	0.07	0.82	1.0	1.0

However, Juxtapose reported the most appropriate matches compared to the results of these two alignment methods, perfectly aligning the networks in every case. This is especially noteworthy given that Juxtapose also did not require any known matches between genes to be provided in terms of external biological information such as sequence similarity and was mainly using network topology to align the networks. Juxtapose was able to align the true matches only using the cosine distance between gene embeddings followed by the Hungarian algorithm to determine the match with the lowest cost.

5.4.2 Alignment of different networks

First, to assess the choice of hyperparameters, we compared the average distance between anchor nodes and random genes to ensure that the selected anchors used to build synthetic structures into the real networks were appropriate for the analyses. We generated a selection of 1000 sets of random genes of equal size to each synthetic structure and compared the sum of the similarities between matched genes in both groups. The distances between the anchor nodes was significantly less than the distances between nodes in the random groups of genes (p-value ¡ 0.001). Therefore, the hyperparameters selected as well as the anchor genes were determined to be appropriate for the following comparisons.

Next, to assess the alignment of different networks, we generated 2 replicate GCNs from subsets of non-overlapping brain and heart samples. As such, these replicates generated similar, but not equivalent, network structures. Each replicate was compared in a pairwise fashion using Juxtapose, and the proportion of correctly matched genes between different networks constructed from the same tissues as well as between altogether different tissues was recorded. The proportion of matches was significantly higher (0.69 and 0.85) when comparing the same tissues vs. when comparing between tissues where the proportion of matches was never more than 0.3. Further, the global similarity values for Juxtapose are shown in Table 5.3. Juxtapose reported global distance scores around 0.3 between tissues; i.e., GCNs that are less similar to each other, for the GCN comparisons made between brain and heart compared to the distances reported for comparisons between GCNs from the same tissue type. Also, the heart GCNs result in global distances that were higher

than the ones reported for brain networks. This likely has to do with the number of edges in the heart networks that form a "hairball" topology. Although the most similar genes tend to be the corresponding genes in the other network, the distance between the genes is much higher.

Similarity measures for IsoRankN and MAGNA++ when all bitscores for matched genes are provided are shown in Table 5.3 and Table 5.4. The proportion of matched nodes did not agree with the similarity between brain and heart networks when using IsoRankN or MAGNA++. The two brain networks were reported as most similar by IsoRankN after the self comparisons, which is reasonable. The next most similar alignment occurred when comparing a brain network to a heart network (0.39). The comparison between the two replicate heart networks is one of the lowest scores (0.34). MAGNA++ had a relatively low percentage of matched nodes between networks. However, MAGNA++ has an S3 score and node score shown in Table 5.4 that reflect the similarity of the networks and it is usually comparable to Juxtapose (but again, MAGNA++ is using bitscores while Juxtapose is not). This score penalizes GCN alignments that map denser network regions to sparser ones or alignments that map sparser network regions to denser areas. However, the proportion of matched nodes remains relatively low and the similarity when comparing heart networks to brain networks is much lower than the scores reported by Juxtapose even though the genes present in these networks and their structures overlap significantly.

Table 5.3: The proportion of genes matched between heart and brain networks compared using IsoRank, MAGNA++, and Juxtapose.

MAGNA++	brain 1	brain 2	heart 1	heart 2	IsoRankN	brain 1	brain 2	heart 1	heart 2	Juxtapose	brain 1	brain 2	heart 1	heart 2
brain 1	1.0	0.03	0.00	0.00	brain 1	1	0.78	0.39	0.35	brain 1	1.0	0.85	0.27	0.29
brain 2		1.0	0.01	0.01	brain 2		1	0.34	0.34	brain 2		1.0	0.30	0.30
heart 1			1.0	0.04	heart 1			1	0.34	heart 1			1.0	0.69
heart 2				1.0	heart 2				1	heart 2				1.0

Table 5.4: The S3 similarity and node score between heart and brain networks compared using MAGNA++, and Juxtapose global cosine distances. Using MAGNA++, a value closer to 1 indicates the networks are more similar and a distance closer to 0 means the networks are more distant. Bitscores were provided to each method and the alpha value was set at 0.5. For the global distance measure, a distance closer to 1 indicates the networks are less similar and a distance closer to 0 means the networks have more similarity.

MAGNA++ S3	brain 1	brain 2	heart 1	heart 2	MAGNA++ NS	brain 1	brain 2	heart 1	heart 2	Juxtapose	brain 1	brain 2	heart 1	heart 2
brain 1	1	0.27	0.16	0.16	brain 1	1.0	0.83	0.55	0.55	brain 1	0.07	0.16	0.30	0.30
brain 2		1	0.15	0.14	brain 2		1.0	0.54	0.52	brain 2		0.07	0.28	0.29
heart 1			1	0.29	heart 1			1.0	0.93	heart 1			0.13	0.22
heart 2				1	heart 2				1.0	heart2				0.11

We also compare MUNK to Juxtapose using the heart GCN replicates. The synthetic networks constructed in Section 5.4.1 were not used to compare to MUNK as there are some characteristics that make them unsuitable input to MUNK. The method has been designed for PPI networks that are sparse, unweighted, and directed. Some of the limitations of MUNK for co-expression networks include the following items.

- Requires networks to be directed
- Does not utilize edge weights
- Removes nodes with a degree less than 2 so their representation will never be learned
- Requires one-to-one orthologs mapping to perform alignment of the networks
- Only analyzes the largest connected component, so if there are two connected components, it will only take into account the largest one and the rest of the genes are lost before the comparison is made

We take the upper triangular correlation matrix of the heart and brain replicate GCNs to form a directed version of the network and remove the weights from the edges. The largest connected component of the networks were 154 and 153 nodes, respectively so these were the components used to make the alignment. Since MUNK uses a linear mapping, it was capable of producing an almost exact match between duplicate heart networks (98.7%). However, MUNK was only able to align 1 (<1%) of the genes successfully between the heart and brain networks where Juxtapose was able to align roughly 30% of the genes between these networks. This may be due to the ability of Juxtapose to consider the edge weights as well as the undirected nature of the networks, allowing the random walks to pass through an edge in either direction and learning more informative embedding in the context of GCN comparison compared to PPI comparison. In this way, Juxtapose can identify genes with similar connectivity in different networks more successfully in the context of GCN comparison.

Figure 5.3 shows the result of biclustering the cosine distance matrix comparing the heart and brain networks with spectral biclustering. The brain GCNs had the most similarity overall, with the most conserved bicluster containing the lowest cosine distances containing genes that were mostly from the Alzheimer disease and Parkinson's disease KEGG pathways (96% of the genes in the top left bicluster were part of the brain disease pathways, and the next most conserved module contained 25% of the heart-related genes). The heart GCNs, on the other hand, resulted in the most conserved bicluster containing a large portion of genes from the KEGG heart-related pathways (30% of the genes were from the heart KEGG pathways in the most conserved bicluster in the bottom right corner, and 26% of these genes were present in the next most conserved bicluster). Furthermore, when comparing one of the heart GCNs and a brain GCN, the genes that were closest to each other based on cosine similarity were mostly from the Alzheimer disease and Parkinson's disease KEGG pathways in both the brain and heart network as opposed to genes specific to the heart KEGG pathways. In the bicluster with the smallest distances between genes, the heart GCN had 78% of the genes from the brain KEGG pathways and 19% from genes shared in both the heart and brain KEGG pathways. Only 3% were strictly from the heart KEGG pathways.

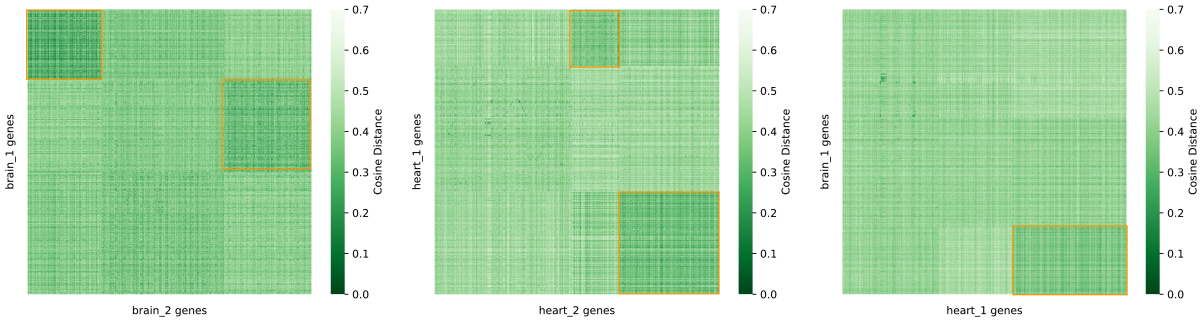


Figure 5.3: Biclustering results for the cosine distance matrix for one replicate of the heart GCNs and one replicate of the brain GCNs. Dark green indicates less distance between nodes while light green or white indicates nodes are more distant from each other. The orange boxes indicate groups of genes that have lower local cosine distances, indicating more conservation between these genes in the GCNs being compared.

5.4.3 Prefrontal cortex multi-species

Finally, we utilized Juxtapose to compare large, real GCNs from different species. Bozek et al. observed an acceleration of metabolite concentration differences among tissues that were confirmed by expression-level differences in corresponding genes in the prefrontal cortex of the brain and in skeletal muscle [3]. They also predicted that these rapid changes might reflect parallel mechanisms in human evolution. We attempt to find evidence of differences in the gene expression regulating the metabolome by constructing GCNs from the data generated by Bozek et al. and comparing the networks using Juxtapose. In the biclusters, we identified groups of genes that were far apart—and thus possible candidates for adaptation with respect to mammalian brain metabolomics—between the species which we selected and performed over-representation analysis to identify modules with enriched KEGG pathways.

Juxtapose was able to identify multiple biclusters with enrichment for KEGG pathways associated with metabolism. Many of these biclusters also had relatively low cosine distances between human and the other three species, suggesting differences in topology in portions of the networks. For example, when comparing human and chimpanzee GCNs biclusters with cosine distances over 0.5, i.e. relatively distant, had enriched terms including Amphetamine addiction, Dopaminergic synapse, and Thyroid hormone signaling pathway, which were reported in the paper by Bozek et al. and include genes that are important regulators of growth, development and metabolism. The biclusters that showed the most difference between these species also included enrichment for choline metabolism. All the compared species had biclusters with high cosine distances (indicating less similarities in these parts of the networks from a topological perspective) in biclusters containing KEGG pathways including Valine, leucine and isoleucine degradation, Inositol phosphate metabolism, Tryptophan metabolism, Pyruvate metabolism, beta-Alanine metabolism, and Propanoate metabolism, beta-Alanine metabolism, some of which were also identified by Bozek et al. Glutamatergic synapse and Aminoacyl-tRNA biosynthesis were terms enriched in a bicluster that was slightly more similar i.e., had lower cosine distances between these two species. These results are in support of human-specific

metabolic divergence as found by Bozek et al [3]. From the global cosine distance score, macaque and mouse were the most distant from human with global cosine distances of 0.41 and 0.40, respectively and human was the most similar to chimpanzee with a global cosine distance of 0.34. These results are presented in Table 5.5. As the global cosine distances are relatively low for all of the species, these results suggest that there is a lot of conserved portions of the networks as well. These global cosine distance results are also supported by the WGCNA results described below. This observation supports that the global cosine distance scores reported by Juxtapose can also be informative when analysing large GCNs. The biclustering enrichment analysis results for each species are presented in the Supplementary Materials. Below, we describe the similarities and differences between the results discovered using Juxtapose and the well-established GCN analysis tool WGCNA.

Table 5.5: Global cosine distances reported by Juxtapose when comparing prefrontal cortex GCNs from human, chimpanzee, macaque, and mouse. For the global distance measure, a distance closer to 1 indicates the networks are less similar and a distance closer to 0 means the networks have more similarity.

	human	chimpanzee	macaque	mouse
human	0	0.34	0.41	0.40
chimpanzee		0	0.36	0.36
macaque			0	0.35
mouse				0

The results of the WGCNA analyses are presented in Figure C.1 and C.2 of the Supplementary Materials. The hierarchical clustering results showed similar patterns for human and chimpanzee gene modules with the macaque clustering appearing the most distinct with one cluster containing a large proportion of the genes. Mouse, on the other hand, had more visual similarity with the human and chimpanzee clustering results. However, the Zsummary scores were relatively low in human versus mouse compared to human versus the other two species (chimpanzee and macaque). The mouse transcriptome being the most distinct from the other three species agrees with the original publication, which concluded that the human metabolome underwent greater change in a shorter period of time than the mouse metabolome did over the 130 million years separating mice from the common ancestor of humans, chimpanzees, and macaques [3]. Mouse is also the most phylogenetically distant from human among these species. We selected modules that showed little to no evidence of preservation (Zsummary < 2) and performed over-representation analysis to identify modules with enriched KEGG pathways. The cyan and pale turquiose modules were the only modules with low preservation that were enriched for any KEGG pathways. Pancreatic secretion, Protein export, Longevity regulating pathway, and Oocyte meiosis, were enriched in the pale turquoise module while the cyan module was enriched with Legionellosis. Of these enriched terms, Pancreatic secretion and Oocyte meiosis are the only enriched terms in the low preservation modules that overlap with the terms reported by Bozek et al. as enriched in the human–specific concentration profiles in the prefrontal cortex. In fact,

most of the enriched pathways show up in the highly conserved modules such as the turquoise module, which includes enriched terms such as Amphetamine addiction, Cocaine addiction, Dopaminergic synapse, Chemokine signaling pathway, Aminoacyl-tRNA biosynthesis that were identified in the original publication. This suggests that although the expression levels of the genes in these clusters may be quite different, they have not changed as much in terms of their co-expression with other genes.

5.5 Discussion

This paper introduced Juxtapose, a tool for comparing the topology of GCNs utilizing a gene embedding approach. One benefit of using Juxtapose as a means of comparing networks is that no knowledge is required about the genes themselves from a biological perspective in order to make a relatively good alignment compared to other alignment methods. Using this embedding method, it is easy to identify not only the best matches with a gene in a corresponding network, but also observe the similarity of a gene to all other genes in the network as well with the local cosine distances. In this way, it is possible to identify areas in the networks that are unambiguous matches (highly conserved) vs. more ambiguous matches (good matches to many genes). It also allows for orthologs that are not strictly one-to-one or functional orthologs to be analyzed to get a more complete picture of the similarities and differences between GCNs, which is a particularly attractive feature for evolutionary studies.

Juxtapose appears to outperform existing alignment-based methods for identifying similar nodes/genes. Indeed, even when aligning artificial networks with unique structures, the typical alignment-based methods performed poorly without prior knowledge of gene similarity. This makes these methods not as informative for aligning co-expression networks. Our method was able to identify the known matches between the identical networks without knowledge of gene similarity. We show that the score of the known matches is also the minimum score one can get by employing the Hungarian algorithm for making the global alignment of the nodes in each network. Therefore, Juxtapose was able to outperform these alignment methods even though its intended purpose is not necessarily to align corresponding nodes in the graphs, but to obtain a measure of similarity between all genes being compared between GCNs.

Juxtapose also outperforms MAGNA++, IsoRankN, and MUNK for aligning different GCNs to one another. MAGNA++ and IsoRankN are only able to achieve comparable results to Juxtapose when they are provided knowledge of the similarity between genes based on biological information such as bitscores. Juxtapose has no such requirement. MUNK also requires some knowledge of orthologs for landmark selection; however, the requirements that likely cause the method to perform more poorly on GCNs compared to PPI networks are that it requires a directed network as input, and it cannot utilize the edge weights to make the alignment. Furthermore, MUNK only operates on the largest connected component of the graph, which may lead to the similarities between some genes not being calculated. Juxtapose, on the other hand, is able to report both local and global distances or similarities between all genes in a GCN.

Another benefit of the proposed methodology is that since it relies on probabilistic walks through the co-expression networks, differences at the level of gene expression or correlation across species do not require normalization across the networks being compared. Normalization tends to be a significant challenge in gene expression analysis, especially when the data has been sequenced in different batches, labs, etc. However, needing to apply multiple types of normalization can actually obscure real signal in the data as none of them work perfectly [24]. Furthermore, there may be unknown factors that require normalization that are missed [29]. Investigating how the length of the walks influences the gene representations in a network embedding, and if an exact walk length can be suggested depending on the number of nodes and edges in a network, is recommended as future research.

Methods such as IsoRank and IsoRankN have been utilized for comparing co-expression networks, but as they are not originally designed for analysing these types of networks. Therefore, assumptions have to be made about the data that may limit the analyses of GCNs. MUNK also has assumptions that may limit the analysis of GCNs, so although these methods may work well for analysing PPI networks, more methods that are specifically designed for comparing GCNs are required. Juxtapose is much more adaptive for networks that require weights on the edges compared to many alignment strategies originally designed for PPI networks.

We also demonstrated that the local cosine distances comparing genes from different GCNs is biologically informative. The biclustering results of the heart and brain GCNs revealed that there was more conservation observed between the genes from the brain-related KEGG pathways. The genes of the heart pathways were more conserved in the heart GCN compared to the brain GCN. This also supports the hypothesis that more conservation would be observed in genes important for regulating processes in the brain, as is supported or suggested in the literature [5].

Ultimately, the goal is to utilize this method in order to compare networks constructed from homologous samples in different species, so we also applied this method to networks constructed from gene expression data from different species. These results could indicate genes that show more evidence of constraint or adaptation between the networks compared. The biclustering results analysing the local cosine distances between human, chimpanzee, macaque, and mouse identified modules of genes that contained enriched KEGG pathways related to metabolism. Also, these modules of genes tended to have high cosine distances, suggesting that these portions of the GCNs across species were less conserved. Juxtapose was also able to identify more terms specifically related to different metabolites compared to both WGCNA and the hierarchical clustering performed by Bozek et al. [3] while also having results that agreed with the observations made by Bozek et al. as well. These results show that Juxtapose can produce results that complement WGCNA results while making it easier to determine the distances or similarities for all pairwise comparisons between modules of genes.

One limitation is the need to confirm that the anchor nodes and synthetic pieces of the network are spread out in different areas. More work can be done to investigate the impact of anchor node selection

from homeostasis-related genes. This could be approached by comparing embeddings using larger subsets of anchor genes as well as genes that have varying relationships across species to determine the impact of anchor gene selection. Future work may also include exploring different updates to the loss function. If they were incorporated directly into the gene expression data, this would not be an issue and will be a goal of future research. There are also newer state-of-the-art embedding strategies in NLP that use Transformers available that can be adapted to embed networks such as BERT [8], ELMO [32] etc. It would be interesting to apply these context dependent methods in future research, particularly with biological networks that have direction to their edges.

5.6 Conclusion

Gene co-expression networks are not easily comparable due to their complex structure. In this paper, we proposed a python-based tool and similarity measures that can be utilized for comparative co-expression network analyses. A word embedding strategy commonly used in natural language processing was adapted and utilized in order to generate gene embeddings based on walks made throughout the gene co-expression networks.

The utility of Juxtapose was demonstrated in scenarios such as comparisons between species and tissues. Synthesized datasets, RNA-seq datasets from GTEx, and a multi-species experiment of prefrontal cortex samples from the Gene Expression Omnibus (GEO) were used to demonstrate its ability to embed the nodes of synthetic structures in the networks consistently while also generating biologically informative results in real networks. Furthermore, Juxtapose is able to successfully align GCNs without relying on known biological similarities and enables post-hoc analyses using biological parameters, such as orthology of genes, or conserved or variable pathways.

References

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. How genomes evolve. In *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[2] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.

[3] Katarzyna Bozek, Yuning Wei, Zheng Yan, Xiling Liu, Jieyi Xiong, Masahiro Sugimoto, Masaru Tomita, Svante Pääbo, Raik Pieszek, Chet C Sherwood, et al. Exceptional evolutionary divergence of human muscle and brain metabolomes parallels human cognitive and physical uniqueness. *PLoS Biology*, 12(5):e1001871, 2014.

[4] Alvis Brazma, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Jaak Vilo, Niran Abeygunawardena, Ele Holloway, Misha Kapushesky, Patrick Kemmeren, Gonzalo Garcia Lara, et al. ArrayExpress: A public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, 31(1):68–71, 2003.

[5] Esther T Chan, Gerald T Quon, Gordon Chua, Tomas Babak, Miles Trochesset, Ralph A Zirngibl, Jane Aubin, Michael JH Ratcliffe, Andrew Wilde, Michael Brudno, et al. Conservation of core gene expression in vertebrate tissues. *Journal of Biology*, 8(3):1–17, 2009.

[6] Jonghwan Choi, Ilhwan Oh, Sangmin Seo, and Jaegyoon Ahn. G2Vec: Distributed gene representations for identification of cancer prognostic genes. *Scientific Reports*, 8(1):13729, 2018.

[7] Chi Tung Choy, Chi Hang Wong, and Stephen Lam Chan. Embedding of genes using cancer gene expression data: biological relevance and potential application on biomarker discovery. *Frontiers in Genetics*, 9:682, 2018.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[9] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[10] Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, 20(1):82, 2019.

[11] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[12] Jason Fan, Anthony Cannistra, Inbar Fried, Tim Lim, Thomas Schaffner, Mark Crovella, Benjamin Hescott, and Mark DM Leiserson. Functional protein representations from biological networks enable diverse cross-species inference. *Nucleic Acids Research*, 47(9):e51–e51, 2019.

[13] Stephen P Ficklin and F Alex Feltus. Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. *Plant Physiology*, 156(3):1244–1256, 2011.

[14] Stephen L France, J Douglas Carroll, and Hui Xiong. Distance metrics for high dimensional nearest neighborhood recovery: Compression and normalization. *Information Sciences*, 184(1):92–110, 2012.

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT press, 2016.

[16] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864. ACM, 2016.

[17] Pietro Hiram Guzzi and Tijana Milenković. Survey of local and global biological network alignment: the need to reconcile the two sides of the same coin. *Briefings in Bioinformatics*, 19(3):472–481, 2018.

[18] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[19] Yuval Kluger, Ronen Basri, Joseph T Chang, and Mark Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research*, 13(4):703–716, 2003.

[20] Harold W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[21] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, 2008.

[22] Charity W Law, Monther Alhamdoosh, Shian Su, Xueyi Dong, Luyi Tian, Gordon K Smyth, and Matthew E Ritchie. RNA-seq analysis is easy as 1-2-3 with limma, glimma and edgeR. *F1000Research*, 5, 2016.

[23] Chung-Shou Liao, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25(12):i253–i258, 2009.

[24] Farhad Maleki and Anthony Kusalik. A synthetic kinome microarray data generator. *Microarrays*, 4(4):432–453, 2015.

[25] Noël Malod-Dognin, Kristina Ban, and Nataša Pržulj. Unified alignment of protein-protein interaction networks. *Scientific Reports*, 7(1):1–11, 2017.

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[27] K Muandet, K Fukumizu, B Sriperumbudur, and B Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–144, 2017.

[28] Mariana F Nery, Brunno Borges, Aline C Dragalzew, and Tiana Kohlsdorf. Selection on different genes with equivalent functions: the convergence story told by hox genes along the evolution of aquatic mammalian lineages. *BMC Evolutionary Biology*, 16(1):113, 2016.

[29] Vegard Nygaard, Einar Andreas Rødland, and Eivind Hovig. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, 17(1):29–39, 2016.

[30] Tun-Wen Pai, Kuan-Hung Li, Cing-Han Yang, Chin-Hwa Hu, Han-Jia Lin, Wen-Der Wang, and Yet-Ran Chen. Multiple model species selection for transcriptomics analysis of non-model organisms. *BMC Bioinformatics*, 19(9):284, 2018.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, 2017.

[33] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

[34] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[35] Elise AR Serin, Harm Nijveen, Henk WM Hilhorst, and Wilco Ligterink. Learning from co-expression networks: possibilities and challenges. *Frontiers in Plant Science*, 7:444, 2016.

[36] Rohit Singh, Jinbo Xu, and Bonnie Berger. Pairwise global alignment of protein interaction networks by matching neighborhood topology. In *Annual International Conference on Research in Computational Molecular Biology*, pages 16–31. Springer, 2007.

[37] Abhijeet R Sonawane, Scott T Weiss, Kimberly Glass, and Amitabh Sharma. Network medicine in the age of biomedical big data. *Frontiers in Genetics*, 10:294, 2019.

[38] Huynh Thanh Trung, Nguyen Thanh Toan, Tong Van Vinh, Hoang Thanh Dat, Duong Chi Thang, Nguyen Quoc Viet Hung, and Abdul Sattar. A comparative study on network alignment techniques. *Expert Systems with Applications*, 140:112883, 2020.

[39] Vipin Vijayan, Vikram Saraph, and T Milenković. MAGNA++: Maximizing accuracy in global network alignment via both node and edge conservation. *Bioinformatics*, 31(14):2409–2411, 2015.

[40] Madeline C Weiss, Martina Preiner, Joana C Xavier, Verena Zimorski, and William F Martin. The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genetics*, 14(8):e1007518, 2018.

[41] Yoseop Yoon, Jeff Klomp, Ines Martin-Martin, Frank Criscione, Eric Calvo, Jose Ribeiro, and Urs Schmidt-Ott. Embryo polarity in moth flies and mosquitoes relies on distinct old genes with localized transcript isoforms. *Elife*, 8:e46711, 2019.

[42] Xinghuo Zeng, Matthew J Nesbitt, Jian Pei, Ke Wang, Ismael A Vergara, and Nansheng Chen. OrthoCluster: a new tool for mining synteny blocks and applications in comparative genomics. In *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*, pages 656–667, 2008.

[43] Bai Zhang, Ye Tian, and Zhen Zhang. Network biology in medicine and beyond. *Circulation: Cardiovascular Genetics*, 7(4):536–547, 2014.

[44] Bing Zhang, Stefan Kirov, and Jay Snoddy. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*, 33(suppl_2):W741–W748, 2005.

Chapter 6

Insights into skeletal cell evolution using Juxtapose

Cell fate decisions are regulated by gene networks, and the knowledge about the evolutionary origin of the gene networks active in skeletal tissues is not well understood. Therefore, a wide-scale gene network comparison across different species would be valuable for understanding the evolution of the skeleton. By identifying global and local similarities in gene expression data across cartilage and osteoblast gene co-expression networks, inferences can be made about the evolution of the skeleton.

This manuscript offers an additional example of an application of Juxtapose introduced in Chapter 5. The methodology from Chapter 5 was utilized in order to compare the gene co-expression networks that can be generated from skeletal tissue RNA-seq data generated by the Eames lab at the University of Saskatchewan. This work will be combined with work being done by other students in the Eames lab.

6.1 Introduction

Understanding the origins of morphological features in organisms is a core challenge in evolutionary biology. This is commonly done by identifying evidence of homology such as through similarities in morphological structures and phylogeny [25]. Another option to study structure origin is to investigate "deep homology", which was first defined by Shubin et al. as "the sharing of the genetic regulatory apparatus used to build morphologically and phylogenetically disparate features" [19]. Identifying conserved gene expression and regulatory relationships during the development of particular organism's structures may identify homology that could not be identified otherwise [22]. The possibility of measuring gene expression using high-throughput next-generation sequencing now offers the potential for identifying a genetic basis of how phenotypic traits have evolved in virtually any non-model organism.

Although cells contain the same DNA, they may differentiate into various types of cells depending on their differentiation program. The fates of these cells are, in part, influenced by coordinated activity of genes. This activity is typically represented as a gene regulatory network (GRN) or a gene co-expression network (GCN). The gene–gene relationships represented in these networks offer information about the homology of organisms alongside other means of establishing and studying homology including morphology, histogenesis, and cell lineages of origin. Development can be constrained by the underlying genetics of an organism [9]. Highly conserved portions of the underlying networks reflect the fundamental circuitry that has been present across long stretches of evolution. Rewiring of these networks can also be used to detect evidence of co-option to generate new anatomical structures [19]. The extent of deep homology across organisms is unclear, and it will be necessary to make many comparisons of the underlying genetic mechanisms of similar and dissimilar structures across diverse organisms to gauge if it is the rule or the exception.

A wide-scale gene network comparison across different species would be valuable for understanding the evolution of the skeleton. The most abundant tissues in vertebrate skeletal tissues are bone, immature cartilage, and mature cartilage. Immature cartilage and mature cartilage differ where immature cartilage will not mineralize, but instead persist over an organism's lifetime, and mature cartilage will mineralize and is typically degraded when replaced by bone [6]. Due to the similarities observed in the functional, embryonic, and histological properties of these tissues, it has been hypothesized that there are core gene networks active across the tissues [8].

The knowledge about the evolutionary origin of the gene networks active in skeletal tissues is not well understood, but there may be some hints as to the origin of bone that can be inferred from their development. Candidate gene approaches that directly test associations between genetic variation within select genes of interest and phenotypes have led to the discovery of critical genes for the regulation of cartilage development [3, 21]. Mechanisms of cartilage development have been widely investigated, but little is understood about how the gene regulatory relationships have assembled throughout evolution [21]. The analysis of specific genes of interest and their expression patterns has suggested that although cartilage in different species

may be regulated using divergent gene regulatory programs, the evolution of cartilage development across organisms has been facilitated by deeply conserved genetic programs [2, 3]. Distantly related species tend to use a remarkably conserved gene activity during embryogenesis [13]. Bone is a unique tissue to vertebrates and may develop through endochondral ossification. The process of endochondral ossification begins with the differentiation of loosely associated cells called mesenchymal cells into chondrocytes, which are the cells that produce cartilage [16]. Chondrocytes may persist or enlarge to become hypertrophic chondrocytes. This is gradually replaced by bone. These mesenchymal cell fates are dictated by gene activity in the skeletal cells. This development of bone is hypothesized to reflect the evolutionary succession of bones [8]. If bone did indeed gradually evolve from cartilage tissues, evidence could likely be found in conserved relationships present between gene expression patterns in both tissue types. Further, we hypothesize that species that diverged earlier in evolutionary history show more similarities between their immature cartilage and bone gene expression patterns. This similarity is due to the establishment of the genetic mechanisms required for bone development being gradually adapted from the underlying genetic mechanisms that were already established in immature cartilage.

Evidence of this co-option of the immature cartilage gene expression network for bone development may be supported by the known molecular contributors to the development of cartilage and bone including Sox9 and Runx2 [12, 15]. Sox9 and Runx2 are candidate transcription factors driving the GRNs responsible for cartilage and bone development, respectively. Sox9 is the earliest indicator of mesenchyme differentiating into chondrocytes producing cartilage [6, 7] while Runx2 is considered a master regulator of bone development [20]. Consistently high levels of Sox9 will commit cells to chondrogenesis to produce cartilage, whereas higher levels of Runx2 will push them toward osteogenesis or bone development [7]. Whether bone develops after immature cartilage depends upon additional transcriptional control by Sox9 or Runx2.

Currently, there is a lot more that could be explored as to the genetic machinery required for bone development. Connecting this back to the evolution of GCNs, the GCNs in immature cartilage may have more similarities with osteoblast GCNs in earlier diverged species where bone appeared. Later diverged species are hypothesized to have more distinct gene regulation in their osteoblast cells compared to immature cartilage [15]. Furthermore, identifying homologous areas in the osteoblast networks of different organisms will improve the understanding of essential gene activity required in order for bone to develop. However, to accomplish this requires a means to make comparisons and quantitatively compare these complex networks.

In this paper, we make use of the embedding methodology proposed in Chapter 5 to explore potential relationships among genes of skeletal tissue and attempt to observe evidence of conservation and adaptation between immature cartilage and osteoblasts. The methodology was utilized in order to compare the co-expression networks that can be generated from skeletal tissue RNA-seq data generated by the Eames lab at the University of Saskatchewan (Amir Ashique, Patsy Gomez-Picos, Jason Nguyen). Section 6.2 provides a description of the data and methodology used to explore the evolution of skeletal tissues and Section 6.3 presents the preliminary results of analyzing these tissues. Interestingly, we find that the global similarity

measure established between networks using gene embedding with Juxtapose shows more conserved gene relationships between the GCNs of the same species than when compared to the GCNs in different species regardless of the time of species divergence. Section 6.4 is used to discuss our findings and Section 6.5 ends the paper with a brief conclusion.

6.2 Materials and Methods

6.2.1 Data

RNA-seq data from 2 skeletal cell types—immature cartilage and osteoblast—available from mouse, chicken, frog, and gar were used to make preliminary inferences about how the relationships between genes have changed through evolutionary history. Each RNA-seq dataset from mouse, chicken, frog, and gar was mapped to Ensembl genome builds GRCm38, GRCg6a, Xenopus_tropicalis_v9.1, and LepOcu1, respectively using STAR 3.5.2 [5]. The raw counts were normalized for each species individually using the Trimmed Means of M values (TMM) from the *edgeR* package [17] where highly expressed genes and those genes that have large variations in their expression values are excluded during the normalization process. Any genes that did not meet thresholds were removed from further analyses.

All co-expression networks were generated using Pearson correlation to establish edge weights between genes of a network. These values were transformed by $0.5+0.5PCC(g_i, g_j)$ where g_i and g_j arbitrary genes in the similarity matrix. These transformed similarities were utilized to generate walks through each network. The number of samples used for each species and tissue type are presented in Table 6.1.

Table 6.1: Number of samples used to construct each skeletal cell GCN.

Species	**Tissue**	**# Samples**
Mouse	IMM	3
Mouse	OST	3
Chicken	IMM	4
Chicken	OST	5
Frog	IMM	3
Frog	OST	3
Spotted Gar	IMM	5
Spotted Gar	OST	5

6.2.2 Network comparison

Once the co-expression networks were generated, the embedding method proposed in Chapter 5 was used to generate gene embeddings for each network. Using the anchoring method from Chapter 5, genes from the

GO term "cellular homeostasis" in Table 5.1 were randomly selected as anchor genes as they are responsible for essential cellular processes, and more likely to be conserved across different species [23]. Ten synthetic structures were constructed as anchor points in the networks of random sizes between 10 and 20 nodes in size, and with 15% of the edges between them weighted as 1 while the remaining edges were weighted as 0.

For each network, 1000 walks of length 30 were generated for each gene. These walks were generated based on the weights along the edges connecting the genes of a network. The walks were then used as training examples to produce an embedding for each network. The parameters used for each embedding are presented in Table D.1 in the Appendix. The resulting vector representations were compared across networks using cosine distance. This distance was used to judge the similarity between genes of different networks. The Hungarian algorithm was used to create a global alignment between each network, and the summation of the distances was used to generate global similarity between the networks. All of the embedding was performed using an AWS spot instance Ubuntu Server 18.04 LTS r5n.8xlarge (32 vCPUs, 256 GiB Memory).

Finally, spectral biclustering of the cosine distance matrices was performed to discover biclusters of genes that shared similar cosine distance patterns. Gene set enrichment analysis (over-representation analysis specifically) was performed on biclusters of genes to determine possible biological processes in which these groups of genes may be participating.

6.3 Results

The global distances reported in Table 6.2 indicate that the GCNs of chicken and frog have the lowest global cosine distances (between 0.43 and 0.45) when comparing networks across species when considering all genes of the network. The global cosine distance between chicken and gar were the next most similar set of GCNs, followed by gar and frog. However, the most similar networks are actually within each species, where the immature cartilage and osteoblast GCNs within the same species are actually the most similar. Mouse and gar had the least similar GCNs when comparing any of the GCNs with global cosine distances over 0.8. The cosine distances comparing the immature cartilage and osteoblast GCNs within mouse resulted in the lowest global cosine distance overall (0.200), and gar had the highest global cosine distance (0.269) for within-species comparisons.

We also selected a subset of 26 genes—Acan, Col2a1, Sox9, Sox5, Sox6, Col9a1, Col9a2, Col9a3, Epyc, Fmod, Comp, Abi3bp, Thbs4, Six1, Matn1, Matn3, Matn4, Col10a1, Runx2, Ihh, Ibsp, Dmp1, Sparc, Col1a1, Col1a2, and Arsi—that have been reported as having significant influence in immature cartilage and osteoblast development [1, 11]. The isolated gene co-expression networks of these select genes for each species and tissue are shown in Figure 6.1 and the global cosine distances between these networks are presented in Table 6.3. The immature cartilage networks visually tend to look more similar to each other compared to the osteoblast networks. However, there are differences in the edge weights between each gene that makes the relationships between genes quite different between the species. The osteoblast GCNs tend to form

Table 6.2: Global cosine distance between immature cartilage (imm) and osteoblast (ost) networks in mouse (mu), chicken (ch), frog (fr), and gar. The closer the values are to 1, the less similar the networks. A distance close to 0 means the networks are similar.

	mu-imm	**mu-ost**	**ch-imm**	**ch-ost**	**fr-imm**	**fr-ost**	**gar-imm**	**gar-ost**
mu-imm	0.000	0.200	0.698	0.664	0.617	0.584	0.823	0.819
mu-ost	-	0.000	0.679	0.662	0.625	0.600	0.804	0.809
ch-imm	-	-	0.000	0.263	0.444	0.437	0.532	0.551
ch-ost	-	-	-	0.000	0.454	0.435	0.547	0.572
fr-imm	-	-	-	-	0.000	0.224	0.604	0.604
fr-ost	-	-	-	-	-	0.000	0.585	0.584
gar-imm	-	-	-	-	-	-	0.000	0.269
gar-ost	-	-	-	-	-	-	-	0.000

Table 6.3: Global cosine distance between immature cartilage (imm) and osteoblast (ost) sub-networks in mouse (mu), chicken (ch), frog (fr), and gar. The closer the values are to 1, the less similar the networks. A distance close to 0 means the networks are similar.

	mu-imm	**mu-ost**	**ch-imm**	**ch-ost**	**fr-imm**	**fr-ost**	**gar-imm**	**gar-ost**
mu-imm	0.000	0.001	0.346	0.345	0.576	0.576	0.691	0.691
mu-ost	-	0.000	0.346	0.345	0.576	0.576	0.691	0.690
ch-imm	-	-	0.000	0.000	0.302	0.300	0.418	0.418
ch-ost	-	-	-	0.000	0.301	0.300	0.418	0.417
fr-imm	-	-	-	-	0.000	0.001	0.146	0.146
fr-ost	-	-	-	-	-	0.000	0.146	0.146
gar-imm	-	-	-	-	-	-	0.000	0.000
gar-ost	-	-	-	-	-	-	-	0.000

modules of positively correlated genes that have negative correlation between each module. However, the gene organization in these modules is also fairly different across species. The mouse osteoblast GCN does not have obvious modules of genes and most genes tend to have a negative or positive correlation with many of the other genes. The underlying expression of these genes is presented in Supplementary Materials.

The cosine similarities presented in Table 6.3 interestingly also suggest similarities between the species that reflect the phylogenetic tree shown in Figure 6.2. Based on global cosine distances, the gar and frog have the lowest global cosine distances between networks followed by the mouse and chicken. The chicken and the frog have the next most similar networks in terms of global cosine distances and so on. Of note was the similarity between the sub-networks in the same species, which was also observed when comparing the whole GCNs. The immature cartilage and osteoblast GCNs were the most similar in one species in comparison to the same GCNs in the other species. The cosine distances comparing the sub-networks within a single species

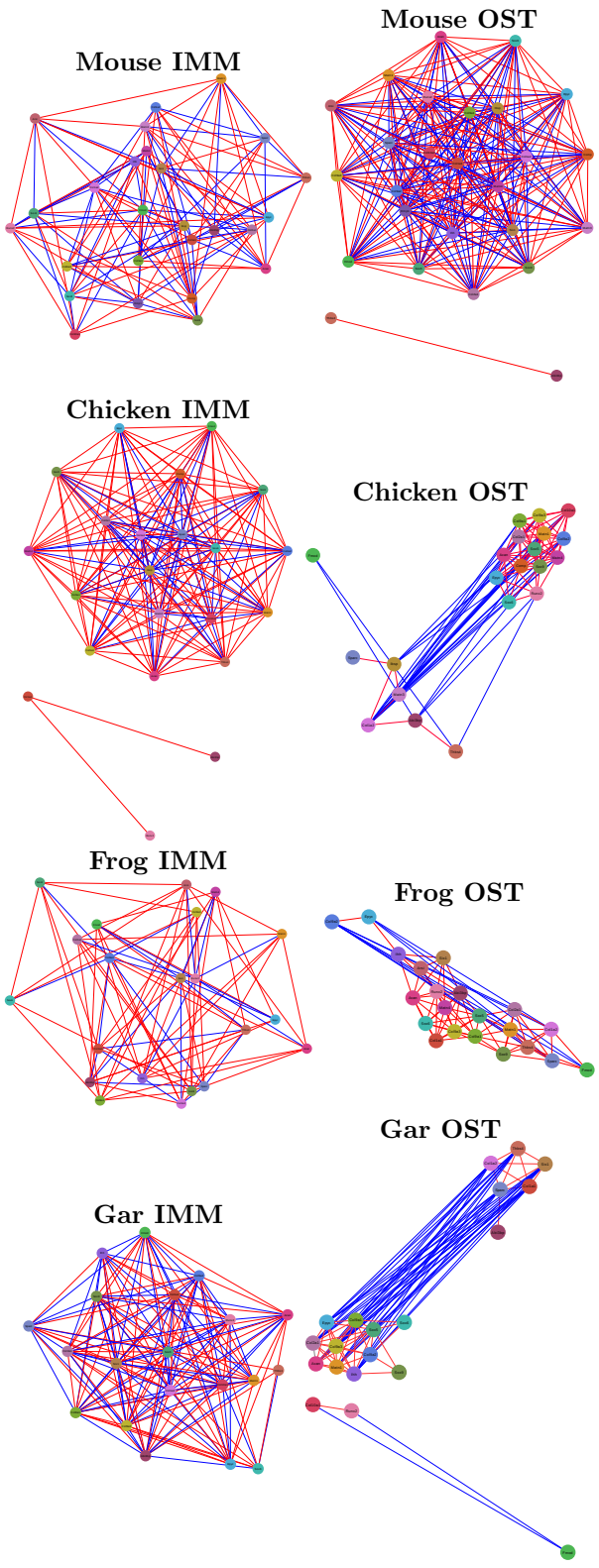


Figure 6.1: Sub-networks of selected genes of interest from immature cartilage (left) and osteoblast (right) GCNs from species shown top to bottom: mouse, chicken, frog, and gar. The red edges indicate edges weighted with a positive correlation and the blue edges indicate edges weighted with a negative correlation. All the colours of the nodes relate to the same genes across all 8 networks, showing the differences between the network organization more clearly.

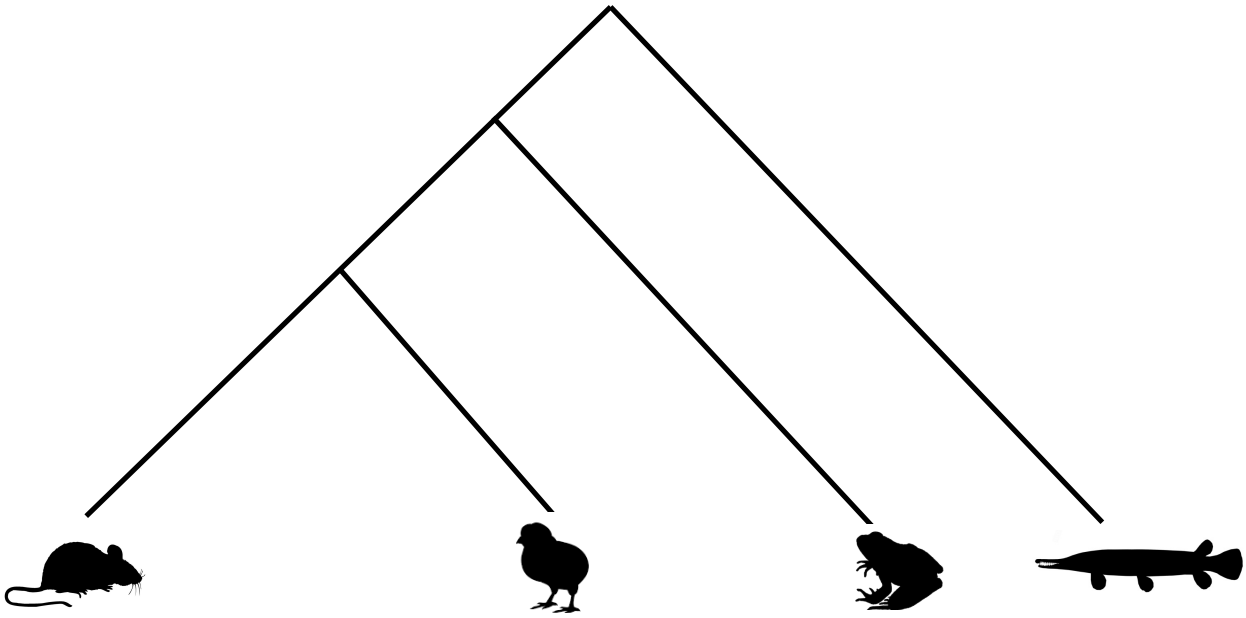


Figure 6.2: A potential phylogenetic tree diagram that shows the evolutionary relationships of mouse, chicken, frog, and gar that have derived from a common ancestor.

stood out as being far more similar in one species versus another.

The biclustering results of the whole osteoblast GCNs revealed similar modules between species that are phylogenetically distant from each other (see Figure D.2). The biclustering results when comparing mouse and gar osteoblast GCNs revealed enriched GO terms such as cell-cell signaling by wnt, where Wnt proteins are a family of secreted glycoproteins that are critical regulators of osteoblast differentiation. ERK1 and ERK2, Notch signaling and activation of MAPK activity were also enriched terms, which have critical roles in the regulation of osteoblastogenesis [14, 24]. This module of genes also had osteoclast differentiation, osteoblast differentiation, positive regulation of osteoblast differentiation, embryonic skeletal system morphogenesis, ossification, and bone development as enriched terms specific to osteoblast development. Genes that were included in the very top left bicluster of Figure D.2 in the Appendix—which was highly conserved with enrichment for terms such as ossification—contained many genes known to be important for bone development including Runx2, Col1a1, Ihh, Spp1, Dlx5, Mmp14, Egr2, and Sp7 [18]. This bicluster also contained a large selection of bone morphogenic protein (BMP) genes, including Bmp1, Bmp2, Bmp3, Bmp4, Bmp6, Bmp7, Bmpr1a, Bmp2k, Bmp8a, and Bmpr2. Other similar modules between the mouse and gar osteoblast GCNs contained enrichment for general cellular processes such as regulation of extrinsic apoptotic signaling pathway, mitotic cell cycle, and translation regulation. Mouse and chicken osteoblast GCNs also had enrichment for general cellular processes, skeletal system development, and the Wnt signaling pathway in the most conserved bicluster in terms of local cosine distances. This bicluster also contained enrichment for specific terms for chondrocyte development and cartilage development involved in endochondral bone morphogenesis. It included genes such as Sox9, Rarg, Trpv4, Col20a1, Matn4, Matn3, Col2a1, Poc1a, Col7a1, Col9a1, Shox2,

Thbs3, Hspg2, Comp, Hoxa11, Matn1, Vwa1, Hoxd11, and Col27a1 involved in these processes. The bicluster with the lowest cosine distances when comparing the osteoblast GCNs of mouse and frog had enrichment for general cellular processes and several terms involving cartilage differentiation and development including genes such as Sox9, Mef2c, Col6a2, Gli3, Col2a1, Vit, Six2, Col11a2, Sfrp2, Anxa2, Col12a1, Loxl2, Tgfbi, Mex3c, Runx2, Matn1, Cytl1, Zbtb16, and Serpinh1. The other highly conserved biclusters had enrichment for general cellular processes.

An interesting observation made from this analysis was that some of the modules of genes that were conserved in mouse and gar had much lower local cosine distances compared to the modules of genes discovered when comparing mouse to frog or chicken, for example. These observations are also shown in Figure D.2. The mouse and gar also had more enriched terms for important processes that aid in the development of osteoblast as well as GO terms specific to skeletal cell development. This means that although the networks globally may be more distinct, there are still modules of genes that may be related to osteoblast development conserved across these phylogenetically distant species. These modules may provide indication as to the portions of the osteoblast network that have been conserved throughout evolutionary history and the less conserved biclusters may indicate the portions of the networks are more subject to change. A representative visualization of each biclustering result are available in Figures D.1, D.2, and D.3 of the Appendix and the enriched terms of each bicluster are available in the Supplementary Materials.

6.4 Discussion

Embedding the genes of the skeletal cell GCNs was performed in order to test the hypothesis that the cartilage and bone GCNs in earlier diverged species are more similar than the cartilage and bone GCNs of species that diverged later in evolutionary history. Our initial observations from measuring global cosine distances indicate that the activity of the genes in these networks in relation to other genes remains relatively consistent between immature cartilage and osteoblast of the same species. This suggests that although the gene expression levels may be changing significantly between the species for these select genes, the influence these genes have either to regulate or act in concert with other genes being expressed are not substantially changing. This would suggest that the levels at which these genes are expressed have far more influence on these phenotypes compared to actual rewiring in the relationships between these genes. These observations do not necessarily support the hypothesis since the whole immature cartilage and osteoblast GCNs in gar were observed to have a higher global cosine distance compared to the global cosine distance between mouse GCNs. However, having more similarity in gene–gene relationships within a single species is not necessarily unexpected; however, if our hypothesis is to be supported, gar would need to show more similarity between its cartilage and osteoblast GCNs. The observation that mouse has the GCNs that are the most similar in terms of global cosine distance does not support this hypothesis.

One possible explanation for similar connections in these GCNs in the same species is that the development

of these tissues are closely linked. Both of these tissues can develop through the process of endochondral ossification. Indeed, it was throughout this process that the RNA-seq samples were collected. As such, it is possible that the relationships between the genes, e.g. the genes regulating the expression of other genes, are the same between the cartilage and osteoblast networks in the same species. The gene expression levels, however, may vary and cause phenotypic differences between the tissue types as they develop. Therefore, utilizing differential expression to study these tissues may be more informative if studying only a single species. However, both differences and similarities were identified between the GCNs when comparing different species so Juxtapose may be more useful in the study of multiple species.

Moreover, one may assume that if gene expression of genes becomes more similar, that the relationships between genes may also be more similar. However, with GCNs this is not necessarily the case. Consider two genes that are normally distributed and not differentially expressed. The correlation of these two genes will approach 0 as the number of samples approach infinity. As such, genes with more similar expression that does not fluctuate tend to appear unrelated in co-expression networks meaning they may be far apart in the network structure. This could be one of the main reasons why genes such as Sox9 may appear less correlated with other genes in a species like gar even though gene expression patterns of this gene becomes similar to other genes expressed in bone compared to the expression patterns observed in mouse.

The global cosine distances also show that although the shape and organization of the networks may at times look different visually in a single species, the correlations in the single species are more preserved than when they are compared across different species. Furthermore, although the topology may look similar in two networks, the connections and organization of the genes may be different. For example, the topology of the sub-networks in Figure 6.1 for chicken immature cartilage and mouse osteoblasts visually appear similar. However, the connections of the corresponding genes between the networks are not very similar. Using force-directed methods for network visualization is a common practice, but is very sensitive to slight differences in the networks [10]. Therefore, visual representations may not be as informative as embedding the genes based on their relationships with surrounding genes of the network.

Reducing the number of genes to a smaller collection of genes was found to better reflect the phylogenetic relationships between mouse, chicken, frog, and gar. This suggests that when a reduced subset of genes known to be involved in a biological process of interest, the similarity between species, at least in terms of cosine distance, may more accurately reflect the phylogenetic relationships between species. The global cosine similarities between GCNs within the same species in this case was the same (0.000). This observation, again, does not indicate that the gar GCNs are more similar to each other compared to the other species under study. However, if the global cosine distance is not rounded, the gar networks are more similar (2.93e-05) than the GCNs within other species (global cosine distances of 3.62e-05 for mouse, 3.45e-05 for chicken, and 3.78e-05 for frog). It is possible that the method may be more applicable to sub-networks as well as being easier to work with compared to running the method on the whole GCN at the same time. In order to focus on sub-networks, there is a challenge to identify the group of genes that should be considered as involved in

a biological process when not much is known about the process.

The biclustering results indicate that although evolutionarily distant species have many differences between the topologies of their skeletal GCNs based on their local cosine distances, there are highly conserved portions of the networks between all the species that have been studied in this work. These conserved biclusters contained enriched GO terms associated with the development of osteoblasts, which suggests that there is a conserved osteoblast GCN across even evolutionarily distant species. Cross-talk between BMP signaling and the signaling pathways of MAPK, Wnt, and Notch, which were enriched pathways in the conserved biclusters between mouse and gar osteoblast GCNs, for example, is also important for osteoblast differentiation and bone formation [4]. A more in-depth study of the conserved modules of genes identified using spectral biclustering is one direction for future research. The conserved biclusters containing genes identified in the literature as important for bone development genes may also include other genes that have not been studied as possible candidate genes important for proper bone development. These results could be useful for future hypothesis generation and identifying genes not previously characterized as important contributors to bone development.

Several caveats of the proposed technique, Juxtapose, could be addressed in future studies with skeletal cells. One limitation of the comparisons between skeletal GCNs is the lack of samples available for each tissue type, which may make it challenging to discover accurate differences and similarities between the GCNs. We showed in Chapter 5 that it is possible to detect differences between tissues using Juxtapose. Whether there is actually a difference between the expression of these genes may be identified with an increase in the number of samples as 200 samples were used to construct the heart and brain GCNs in Chapter 5. One of the networks that may change due to increased samples is the mouse or frog osteoblast network. From Figure 6.1, it was observed that the osteoblast network of mouse is much more highly connected than the other three species which have distinct modules of positively correlated genes. This could be due to only three samples being available for the mouse osteoblast meaning that high correlation may be observed more frequently even when there would be no relationship between a pair of genes if more samples were introduced. These networks may also be supplemented with information available from protein-protein interaction (PPI) networks although this information may not be available for each species, which could bias any downstream comparisons between the networks. Also, the necessity of a network threshold to avoid too many walks being required from a single node in the network leaves the challenge of selecting appropriate thresholds for different networks, in this case from different species. Different thresholding strategies could be explored including rank-based thresholding or soft thresholding. Making it more efficient where the number of walks originating from a gene depends on its degree is another strategy for reducing the number of walks required to represent a GCN.

Lastly, this study compared the GCNs generated using immature cartilage and osteoblast samples, but mature cartilage RNA-seq samples are also available. Mature cartilage is one of the other main types of skeletal tissue that has gene expression that has similarities to both immature cartilage and osteoblasts.

As future research, we suggest comparing the GCNs constructed from mature cartilage as well to gain more insight into how different the relationships between genes may be across species. These insights in conjunction with differential expression information can be used to make evolutionary inferences about the development of these skeletal tissues. Furthermore, this embedding method makes it easier to incorporate more datasets at the level of walk generation as only normalization within each dataset is necessary. As such the incorporation of more datasets from more species may be a possibility in the future.

6.5 Conclusion

This study compared the GCNs generated using immature cartilage and osteoblast samples from RNA-seq samples from mouse, chicken, frog, and gar. We utilized Juxtapose to make comparisons between GCNs from immature cartilage and osteoblasts in these species to detect evidence of conservation and adaptation in the gene relationships of these GCNs. We found that the gene relationships in these GCNs were relatively conserved within a single species compared to the GCNs being compared across different species, which suggests that the expression levels are causing the phenotypic difference between these tissues in a species may be more influenced by relative gene expression levels more so than any rewiring of the networks.

Global cosine distances were used to measure the differences between gene embeddings of immature cartilage and osteoblast in a quantitative manner. Furthermore, biclustering was performed to identify portions of the networks that were more conserved as well as the biological processes that may be associated with these modules of genes. Our observations suggest that there is a substantial amount of conservation between the mouse osteoblast GCN and both the gar immature cartilage and osteoblast GCNs although the global distance measure indicated that overall the networks are relatively different from each other. Genes in these conserved portions may include important genes for skeletal cell development that could be studied in future work.

References

[1] Patrick Aghajanian and Subburaman Mohan. The art of building bone: emerging role of chondrocyte-to-osteoblast transdifferentiation in endochondral ossification. *Bone Research*, 6(1):1–9, 2018.

[2] Thibaut Brunet and Detlev Arendt. Animal evolution: The hard problem of cartilage origins. *Current Biology*, 26(14):R685–R688, 2016.

[3] Maria Cattell, Su Lai, Robert Cerny, and Daniel Meulemans Medeiros. A new mechanistic scenario for the origin and evolution of vertebrate cartilage. *PloS One*, 6(7):e22474, 2011.

[4] Guiqian Chen, Chuxia Deng, and Yi-Ping Li. TGF-β and BMP signaling in osteoblast differentiation and bone formation. *International Journal of Biological Sciences*, 8(2):272, 2012.

[5] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[6] B Frank Eames, Luis De La Fuente, and Jill A Helms. Molecular ontogeny of the skeleton. *Birth Defects Research Part C: Embryo Today: Reviews*, 69(2):93–101, 2003.

[7] B Frank Eames, Paul T Sharpe, and Jill A Helms. Hierarchy revealed in the specification of three skeletal fates by sox9 and runx2. *Developmental Biology*, 274(1):188–200, 2004.

[8] Patsy Gómez-Picos and B Frank Eames. On the evolutionary relationship between chondrocytes and osteoblasts. *Frontiers in Genetics*, 6:297, 2015.

[9] Stephen Jay Gould. The evolutionary biology of constraint. *Daedalus*, pages 39–52, 1980.

[10] Ivan Herman, Guy Melançon, and M Scott Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.

[11] Toshihisa Komori. Regulation of bone development and extracellular matrix protein genes by RUNX2. *Cell and Tissue Research*, 339(1):189, 2010.

[12] Elena Kozhemyakina, Andrew B Lassar, and Elazar Zelzer. A pathway to bone: signaling molecules and transcription factors involved in chondrocyte development and maturation. *Development*, 142(5):817–831, 2015.

[13] Robb Krumlauf. Hox genes in vertebrate development. *Cell*, 78(2):191–201, 1994.

[14] Takehiko Matsushita, Yuk Yu Chan, Aya Kawanami, Gener Balmes, Gary E Landreth, and Shunichi Murakami. Extracellular signal-regulated kinase 1 (ERK1) and ERK2 play essential roles in osteoblast differentiation and in supporting osteoclastogenesis. *Molecular and Cellular Biology*, 29(21):5843–5857, 2009.

[15] Jason KB Nguyen and B Frank Eames. Evolutionary repression of chondrogenic genes in the vertebrate osteoblast. *The FEBS Journal*, 2020.

[16] Ross E Petty and James T Cassidy. Structure and function. In *Textbook of Pediatric Rheumatology*, pages 6–15. Elsevier, 2011.

[17] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[18] Nadeem Samee, Valerie Geoffroy, Caroline Marty, Corinne Schiltz, Maxence Vieux-Rochas, Giovanni Levi, and Marie-Christine de Vernejoul. Dlx5, a positive regulator of osteoblastogenesis, is essential for osteoblast-osteoclast coupling. *The American Journal of Pathology*, 173(3):773–780, 2008.

[19] Neil Shubin, Cliff Tabin, and Sean Carroll. Deep homology and the origins of evolutionary novelty. *Nature*, 457(7231):818–823, 2009.

[20] Gary S Stein, Jane B Lian, Andre J Van Wijnen, Janet L Stein, Martin Montecino, Amjad Javed, Sayyed K Zaidi, Daniel W Young, Je-Yong Choi, and Shirwin M Pockwinse. Runx2 control of organization, assembly and activity of the regulatory machinery for skeletal gene expression. *Oncogene*, 23(24):4315–4329, 2004.

[21] Oscar Tarazona, Leslie Slota, and Martin Cohn. Deep conservation of the genetic program for cartilage development: The mechanism of invertebrate chondrogenesis. *Developmental Biology*, 1(356):246, 2011.

[22] Patrick Tschopp and Clifford J Tabin. Deep homology in the age of next-generation sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1713):20150475, 2017.

[23] Madeline C Weiss, Martina Preiner, Joana C Xavier, Verena Zimorski, and William F Martin. The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genetics*, 14(8):e1007518, 2018.

[24] Stefano Zanotti and Ernesto Canalis. Notch signaling and the skeleton. *Endocrine Reviews*, 37(3):223–253, 2016.

[25] Zhengting Zou and Jianzhi Zhang. Morphological and molecular convergences in mammalian phylogenetics. *Nature Communications*, 7(1):1–9, 2016.

Chapter 7

Contributions, Limitations, and Future Work

In this chapter, I provide a summary of the previous chapters of the thesis, highlighting their contributions and limitations, as well as discuss potential avenues for future research.

7.1 Summary and contributions

In Chapter 2, a comprehensive review of the common ways to compare gene co-expression networks (GCNs) was provided. I classified the current methods used to compare GCNs into alignment-based and alignment-free methods. These techniques were mainly discussed in the context of evolution, highlighting their strengths and shortcomings when comparing GCNs originating from different species, and discussed the challenges that GCN comparative analyses still face. Furthermore, we offered suggestions for possible directions for future research using natural language processing (NLP)-based techniques, which also provided the motivation behind the research presented in this thesis. Considering the insights from the literature review, we designed and conducted several studies to compare GCNs across species and proposed new similarity measures to capture the global and local similarities between GCNs in different species in order to gain insight into their evolutionary relationships.

In Chapter 3, I created a new visualization tool available as an R package (built upon the *ggplot2* package) for the concise visualization of symmetric matrices called *pineplot*. The package provides functions to build symmetric matrices, visualize these matrices as *ggplot2* triangular heat maps and stack them for easy comparison across group, time, phenotype, etc. This is useful for comparisons made across species as well as other variables as there are many symmetric measures used to represent the relationships between genes (Pearson correlation, Spearman correlation, mutual expression, etc.). Not only is this software useful for comparative transcriptomics as highlighted by the cross-species case study presented in the paper, it is also a useful visualization tool for comparing multiple factors such as tissue type and sample size as used in the paper available in Chapter 4. This package is a valuable contribution to the research community at large as many, particularly biological datasets, can be represented using symmetric measures. When more than three variables need to be visualized, *pineplot* offers concise representations of symmetric matrices that facilitate creating a holistic picture of the relationships between these variables across different experimental conditions.

A methodology to evaluate the reproducibility of GCNs across sample sizes and tissue types was proposed in Chapter 4. Rather than relying on simulated datasets or accurate and informative functional annotation for evaluating GCN reproducibility, our generalized method is capable of handling real gene expression datasets,

the potentially large fully connected GCNs resulting from this data, as well as different phenotypes. Our study of four commonly used measures and three different tissue types used to construct GCNs showed that the optimal sample size for generating reproducible networks might be dependent on tissue type and the method used to construct the networks. As this could, in part, be due to differences in the heterogeneity of the data from different phenotypes, the number of samples required to generate reproducible GCNs may also be dependent on the species under study. This study provided insight into how the reproducibility of GCN construction can be impacted. For example, smaller sample sizes may not be as much of a limitation if fully connected networks are used for analysis, such as with WGCNA. The differences between edge weights, even using small sample sizes, also tend to be relatively small. However, the order of the edges may be impacted significantly using small sample sizes. Therefore if a thresholding step is applied to the GCNs for further analysis, it may be necessary to increase the number of samples to produce networks with more consistent ordering of edges.

Chapter 5 contains the results of an NLP-based method I developed to compare genes of different GCNs, where genes are represented as vectors in a lower-dimensional space; i.e., created a gene embedding based on gene expression data. By doing so, distances between genes of interest can be calculated and used as a representation of how similar or dissimilar the genes are when comparing genes within or between co-expression networks. Not only this, I create a joint embedding so that multiple co-expression networks can be compared. This allows for the calculation of global and local similarities between the networks to be calculated more efficiently than a method such as network alignment. The method can be applied to small networks as well as networks containing all expressed genes in an organism. I also demonstrate its utility to compare concordant gene expression patterns between tissues and between species. Our method is more appropriate for comparing GCNs as opposed to using an embedding technique like MUNK (designed for PPI networks), which we demonstrate only works well for directed, unweighted, and relatively sparse networks. It is shown that our method outperforms more traditional methods of comparing GCNs as well.

Finally, in Chapter 6 a wide-scale gene network comparison across the skeletal cell GCNs from different species is performed as an initial study of osteoblast cell evolution. The methodology from Chapter 5 was utilized in order to compare the co-expression networks that can be generated from skeletal tissue RNA-seq data.

This research aimed to develop methods to identify similarities and differences between GCNs in order to observe evidence of change in gene expression relationships using high-throughput expression data—either comparing GCNs originating from different tissues, conditions, or different species to find evidence of evolutionary change. Based on quantitative analyses of GCNs and gene activity from various perspectives, it can be concluded that visually analysing GCNs can offer only so much insight into the actual changes and evidence of conservation present in this type of network. GCNs are often very large and complex structures, and due to their density it can be challenging to determine if the networks contain conserved gene relationships. When these networks have the same edges, as with fully connected networks containing

a reduced set of genes shared across GCNs, it is possible to compare the networks based on the weights of their edges to get a global or local measure of similarity. The analysis of these networks to determine sample size showed such a comparison.

However, two GCNs containing the same set of genes is not common, particularly when making comparisons across species. For example, GCNs may have a threshold to edge weights to remove weaker relationships between genes applied, or they may originate from different species that do not have the same genes present in their genome. In different species, there are challenges with analyzing the network structure of networks when they include genes that do not have a one-to-one ortholog in the other GCN. Perhaps there are also genes that do not have known relationships to genes of another species, or they have multiple orthologs associated with them in other species. Typically, these types of scenarios tend to be ignored in analyses as there are few strategies to incorporate them [14, 21].

I have presented several tools that are useful methods for analysing GCNs. By analyzing changing representations of GCNs from different scenarios, this thesis has shown how the relationships between genes and their expression may be impacted by sample size, different conditions, and how their concordant activity may change in different species. Our methods to compare GCNs may be applied in the future to study changes in gene relationships across evolutionary history with the ability to handle the analyses of non-model organisms where high-throughput gene expression data may be an available type of data that can be collected in a short period. These methods also have the potential to be applied to analyse PPI networks, which can offer complementary information as to similarities and differences in protein activity across different PPI networks. Below the possible limitations and future research inspired by the results of this thesis are discussed in more detail.

7.2 Limitations

In this section, I highlight some of the limitations of the studies conducted in this thesis. It should be noted that a number of the following limitations are not specific to this research, but limitations that are common when utilizing biological datasets that need to be considered when interpreting the results of analyses.

7.2.1 Experimental design and small sample sizes

The design of the experiments being used to generate GCNs is typically an important factor to consider. It has been shown that comparisons made across species can sometimes be confounded by experimental designs that separate species in a way that makes normalization ineffective. If an experiment does not use a random block design, it may be impossible to normalize for different factors since information about each needs to be present in each block. One reason this is difficult in gene expression data from different species is that in many cases, all of the species may not be available at the same time to perform a single sequencing experiment. There are also limitations involving the depth of sequencing as more samples are used. More

species in a single experiment are going to limit depth. Therefore, as in the case of the skeletal cell data, the species have been sequenced separately. This means batch normalization to combine all of the data cannot be performed. However, the embedding method proposed in Chapter 5 bypasses these challenges. In fact, data from other experiments could also be utilized in a transfer learning approach to create more accurate gene representations.

Sample size and tissue type, as demonstrated in Chapter 4, are important factors when trying to generate reproducible gene co-expression networks. I could also extrapolate that species may have a similar impact as tissue type when it comes to generating reproducible networks. This is because different species can have differences in the heterogeneity of their gene expression due to factors such as the species being wild-type versus lab-bred species [9]. Since the application of the embedding method proposed in this thesis was applied to skeletal cell RNA-seq samples where there were only 3-5 samples per group, reproducibility may be limited until more samples can be collected from the same species. The greatest variation may arise from the order of the edges in each network meaning that using a cutoff—as was done before the embedding step of each network—may result in missed edges that are actually important between the genes of a network. As it becomes easier to generate new samples and potentially supplement with other data, these limitations may be mitigated.

The study of optimal sample sizes for the reproducibility of GCNs showed that different datasets may require different numbers of samples in order to generate reproducible GCNs. However, this methodology should be applied to many more phenotypes or conditions in order to confirm if this is a pattern common for many phenotypes. This may also require more sample sizes (more than 50 samples) to be taken into consideration. One benefit of the method proposed in Chapter 4 is that it is a general purpose method that can handle larger networks. However, larger networks, especially if they are fully connected, will limit where evaluations can be performed and may require much more computational power and memory to store and analyse the networks.

7.2.2 Functional annotation

Functional annotation is defined as the process of collecting information about and describing a gene's biological identity, including its aliases, molecular function, biological role(s), location, and its expression domains within an organism. Comparisons or the evaluation of methods across species are sometimes limited by the functional annotation of a gene. For example, gene set enrichment analysis is a common method utilized in order to identify groups of genes potentially related by their functions. However, these methods are known to have serious issues that limit their value in biological evaluations [15–17, 19].

One limitation of many methods used to compare genes across species is that it is often necessary for the same genes to be compared across the species. This means that the genes tend to be limited in some way, such as reducing to only one-to-one orthologs in order to make comparisons easier. Using the embedding strategy presented in Chapter 5, there is no such limitation unless the method of evaluation downstream will

only consider gene set analysis results. The relationship a gene has with the other genes of a network can be calculated, and based on these relationships, a distance may still be calculated between the genes of another network. It is the same principle as getting the similarity between different, unrelated genes of two networks. However, due to the way the embeddings need to be made comparable, i.e. embedded in the same space, it is essential to select appropriate anchor genes for comparing networks. This is reliant on accurate annotation, although it is reliant on far fewer genes having accurate annotation.

Although the embedding strategy for comparing GCNs does not require knowledge of the underlying biology of the genes in the networks, this information is valuable when trying to select appropriate genes as anchor nodes. The limitation of requiring known and appropriate anchor genes means it is necessary that researchers utilizing the tool have background knowledge of appropriate genes to select as anchor genes that are going to share a lot of similar functions and relationships in the compared co-expression networks. If arbitrary genes are selected as anchors, the results of the embedding could differ drastically and will likely not be biologically informative. The focus was on genes annotated with general processes required to keep cells alive in order to mitigate this limitation, but it is not guaranteed that these genes have consistent relationships across all species. As such, it is recommended to utilize something similar to the anchor distance comparisons to random gene selections described in Section 5.4.2 of Chapter 5 when selecting potential anchor genes to try and minimize these errors.

7.2.3 Risk of overfitting or underfitting in gene embedding

Overfitting occurs when our model performs well on our training data, i.e. the walks made through the GCN of interest, but performance decreases when the model is presented with an external validation set (walks it has not been presented with). The word2vec architecture [13, 18] was utilized to train our models since overfitting tends to be difficult [8]. More specifically, I utilize the skip-gram architecture since it tends to be less sensitive than continuous bag of words (CBOW) to overfit frequent words, because even if frequent words are presented more times than rare words during training, they still appear individually, while CBOW is prone to overfit frequent words because they appear several times along with the same context.

In the context of embedding GCNs from Chapter 5, overfitting may occur if the dangling structures of two networks being compared are preferably aligned while disregarding other sections of the networks. This can be a difficult scenario to identify without extensive biological evaluation to determine if the alignment of other parts of the network make sense. This is even more challenging when little has been characterized about the networks. it is difficult to determine if a gene alignment makes sense when the function of the genes is unknown. To mitigate this, we rely on very few dangling structures in each network so the alignment does not overwhelm the real portion of the networks. However, if the alignment of the dangling structure was incorporated into the loss function to encourage its alignment in the future, this challenge would need to be considered further.

Underfitting, on the other hand, may occur if enough walks cannot be generated for a node to accurately

represent it and its relation to all other nodes of the network [7]. In this case the embedding created is not informative and requires more training data to make a better representation. However, knowing when more walks, and how many walks, are ultimately required to make a good representation for each gene is challenging especially with larger networks. I preferably take advantage of our ability to make more training examples as opposed to increasing training iterations using the same walks. Underfitting should be relatively easy to identify with the anchor nodes and dangling structures in the network being used for validation. However, adding more training data can be a challenge if underfitting does occur when dealing with large networks. A large amount of memory is currently required for this step in our embedding method as more walks are generated for training the word2vec model.

In machine learning, hyperparameter tuning is the process of selecting the optimal values for hyperparameters for a learning algorithm [2]. This could include the manipulation of constraints, weights, or learning rates to generalize to different data patterns as well as further minimize the results of the loss function. Tuning hyperparameters such as the number of iterations for the embedding strategy is done in order to prevent issues such as underfitting and overfitting. However, when dealing with large networks, hyperparameter tuning can become a significant bottleneck as well. One strategy may be to use random search [2], which searches the specified subset of hyperparameters randomly instead of exhaustively. The major benefit of this approach is decreased processing time. However, finding the optimal combination of hyperparameters is not guaranteed.

7.3 Future directions

In this section, I emphasize some of the ideas for future research proposed throughout this thesis and provide suggestions for making improvements to GCN analysis.

7.3.1 Systematic evaluation of GCN analysis methods

As mentioned in Chapter 4, a systematic evaluation of network comparison methods has not been performed in the context of GCNs. A large-scale evaluation of these techniques would be useful for identifying promising techniques that can be utilized for comparing GCNs. Since it has been reported that common gene regulatory network construction strategies do not perform as well as simple correlation-based strategies [1], I did not consider evaluating such methods based on reproducibility. However, if they were to be evaluated from the perspective of GCN comparison and not reproducibility, the results from a different type of gene network, such as gene regulatory networks, may outperform correlation-based methods. Just because correlation-based methods generate more reproducible networks using smaller sample sizes does not indicate that they are capturing informative biological relationships. Furthermore, gene regulatory networks have direction to their edges, which may be more biologically accurate. It may be interesting to evaluate these strategies and investigate how much information is truly being captured by a gene co-expression network. However, this

requires datasets with known gene relationships, which is probably the major challenge to performing these evaluations. Utilizing time-series data may also increase the information captured by co-expression networks. However, these datasets tend to be quite small in comparison to steady state datasets so they are not typically used for evaluation [5].

7.3.2 Exploration of GCN thresholding strategies

The application of a cut-off to reduce the number of edges in co-expression networks can be done in several ways. One way is to apply a cut-off x that is applied to remove any edge that does not make this cut-off. The cut-off may also be rank-based, where the top $x\%$ of edges is retained for each gene in the network, regardless of the weights of the edges [3]. Or no cut-off may be applied at all (soft thresholding) [12]. In future works, changing the way that these networks are thresholded may be done to determine how the embedding method responds to different thresholding techniques. Depending on the type of thresholding strategy utilized, gene representations may change drastically and at this time it is not known how changing the thresholding will impact the biological insight that can be obtained from using these types of methods. Therefore, I suggest investigating different thresholding techniques to see if one option is best when utilizing embedding techniques to generate gene representations.

Also, investigating network thresholding strategies also relates to the generation of reproducible GCNs. Edge weights that indicate strong relationships between genes may tend to remain more consistent across subnetworks compared to edge weights that indicate weaker relationships. Therefore, certain thresholds may improve network reproducibility when utilizing smaller sample sizes to construct GCNs. As future research, it would be of interest to investigate if there are differences in how consistent edge weights are across GCNs based on their strength/value. Other means of preprocessing RNA-seq datasets, including the use of smoothing techniques [20] could also be attempted to remove more technical noise while trying to preserve biological heterogeneity and improve GCN construction when using smaller sample sizes.

7.3.3 Phylogenetic tree construction

Network alignment has been utilized to estimate species trees in comparison to sequence alignment strategies using the patristic distance to compare the constructed phylogenetic trees [11]. The patristic distance is the sum of the lengths of the branches that link two nodes in a tree, where those nodes are typically terminal nodes that represent extant genes or species. A matrix of patristic distances calculated from a tree for all pairs of genes or species summarizes the genetic change, or phylogenetic change, represented in the tree [6]. Using network alignment, this distance has been calculated using both the similarities between species that are measured by edge scores and the topology of the phylogenetic tree. This suggests that using an embedding strategy, these distances could also be calculated. In order to determine if this is possible using embedding global similarities, more data from different species is required to determine if this pattern is common among species in various portions of a phylogenetic tree. Calculating these distances using an embedding strategy

with GCNs could also provide insight as to how fine-grained the comparison could be; i.e., how similar can the species be before using gene expression and embedding is unable to construct accurate relationships between the species?

7.3.4 Catering loss functions to GCN comparison

Like many other machine learning techniques, word2vec uses gradient descent to minimize the cross-entropy loss over the entire corpus; that is, the probability of predicting the wrong gene. The loss function is the quantity to minimize, given our training example, where we want to maximize the probability that our model predicts the target gene given our context gene. Changing the loss function can be done to improve the alignment of genes in different networks. Currently, the loss function used by our embedding strategy is Negative Sampling proposed by Mikolov et al. [18], which focuses on learning a high-quality word embedding rather than modeling the word distribution in natural language. However, there are many options when selecting a loss function or embedding strategy to create a representation of the genes. Other potential loss functions that may be applicable to GCNs include Margin Ranking Loss and Hinge Loss [10].

7.3.5 Transformers for context-specific embedding

The main problem of word2vec is that it provides a single representation for a word (or gene) that is the same regardless of context. So words with several different meanings will end up with a representation which is an average of the meanings and not represent any one well. It is hypothesized that using these methods would not improve the results using co-expression networks as they are undirected graphs. This means that regardless of the order the interactions happen, this information is not retained in a co-expression network; i.e., no context. Therefore, context-based methods likely would provide much improvement over word2vec. However, context-based methods could be applied for other networks where interactions are more defined than co-expression networks.

BERT (Bidirectional Encoder Representations from Transformers) uses the bidirectional training of Transformer [4], a popular attention model, for language modelling. This is in contrast to previous efforts that looked at a text sequence either from left-to-right or combined left-to-right and right-to-left training such as word2vec. BERT, on the other hand, is able to use the entire context for prediction and not only the left context. One thing BERT demonstrates is that by encoding the context of a given word, by including information about preceding and succeeding words in the vector that represents a given instance of a word, much better results could be obtained in natural language processing tasks. Although not much use for co-expression networks being that their edges lack direction, other biological networks do have direction, such as gene regulatory networks. This is an area of future work where context specific embedding methods could be valuable as it is important to know the order in which genes are able to interact with each other.

References

[1] Sara Ballouz, Wim Verleyen, and Jesse Gillis. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*, 31(13):2123–2130, 2015.

[2] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

[3] Olgert Denas, Richard Sandstrom, Yong Cheng, Kathryn Beal, Javier Herrero, Ross C Hardison, and James Taylor. Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution. *BMC Genomics*, 16(1):1–9, 2015.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Jason Ernst, Gerard J Nau, and Ziv Bar-Joseph. Clustering short time series gene expression data. *Bioinformatics*, 21(suppl_1):i159–i168, 2005.

[6] Mathieu Fourment and Mark J Gibbs. PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. *BMC Evolutionary Biology*, 6(1):1–5, 2006.

[7] A. Ghasemian, H. Hosseinmardi, and A. Clauset. Evaluating overfit and underfit in models of network community structure. *IEEE Transactions on Knowledge and Data Engineering*, 32(9):1722–1735, 2020.

[8] Yoav Goldberg and Omer Levy. word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[9] Tzachi Hagai, Xi Chen, Ricardo J Miragaia, Raghd Rostom, Tomás Gomes, Natalia Kunowska, Johan Henriksson, Jong-Eun Park, Valentina Proserpio, Giacomo Donati, et al. Gene expression variability across cells and species shapes innate immunity. *Nature*, 563(7730):197–202, 2018.

[10] Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. *arXiv preprint arXiv:1702.05659*, 2017.

[11] Oleksii Kuchaiev, Tijana Milenković, Vesna Memišević, Wayne Hayes, and Nataša Pržulj. Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society Interface*, page rsif20100063, 2010.

[12] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, 2008.

[13] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.

[14] John Lee, Manthan Shah, Sara Ballouz, Megan Crow, and Jesse Gillis. CoCoCoNet: conserved and comparative co-expression across a diverse set of species. *Nucleic Acids Research*, 48(W1):W566–W571, 05 2020.

[15] Farhad Maleki, Katie Ovens, Ian McQuillan, and Anthony J Kusalik. Size matters: how sample size affects the reproducibility and specificity of gene set analysis. *Human Genomics*, 13(1):42, 2019.

[16] Farhad Maleki, Katie Ovens, Ian McQuillan, Elham Rezaei, Alan M Rosenberg, and Anthony J Kusalik. Gene set databases: A fountain of knowledge or a siren call? In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 269–278, 2019.

[17] Farhad Maleki, Katie L Ovens, Daniel J Hogan, Elham Rezaei, Alan M Rosenberg, and Anthony J Kusalik. Measuring consistency among gene set analysis methods: A systematic study. *Journal of Bioinformatics and Computational Biology*, 17(5):1940010–1940010, 2019.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[19] Pashupati Mishra, Petri Törönen, Yrjö Leino, and Liisa Holm. Gene set analysis: limitations in popular existing methods and proposed improvements. *Bioinformatics*, 30(19):2747–2756, 2014.

[20] Florian Wagner, Yun Yan, and Itai Yanai. K-nearest neighbor smoothing for high-throughput single-cell RNA-seq data. *BioRxiv*, page 217737, 2017.

[21] Koon-Kiu Yan, Daifeng Wang, Joel Rozowsky, Henry Zheng, Chao Cheng, and Mark Gerstein. OrthoClust: an orthology-based network framework for clustering data across multiple species. *Genome Biology*, 15(8):R100, 2014.

Appendix A

Supplementary material for Chapter 3

Table A.1: Sample IDs per species

Macaque	Mouse	Chicken
SRR306777	SRR306757	SRR594503
SRR306778	SRR306758	SRR594509
SRR306784	SRR306769	SRR594512
SRR306785	SRR306770	SRR306710
SRR306786	SRR306772	SRR306711
SRR306787	SRR306773	SRR306716
SRR306788	SRR594393	SRR306717
SRR594446	SRR594396	SRR306718
SRR594449	SRR594397	SRR306719
SRR594450	SRR594410	SRR594500
SRR594464	SRR594413	SRR594513
SRR594467	SRR594414	SRR594521
SRR594468	-	-

Table A.2: Variables in the UCI liver disease dataset visualized for case study.

Abreviation	Description
Age	Age of patient in years.
Sex	Sex of the patient.
Alkophos	Alkaline phosphatase level.
Sgot	Aspartate aminotransferase level.
Sgpt	Alanine aminotransferase level.
Tbil	Total Bilirubin. Bilirubin comes from the breakdown of red blood cells and is excreted by the liver.
Dbil	Direct Bilirubin: unconjugated bilirubin measurement.
Proteins	Total Proteins
Albumin	Albumin is made by the liver and binds to calcium, hormones, vitamins and drugs and carries them through the bloodstream concentration.
AG ratio	Albumin to globulin (A/G) ratio.

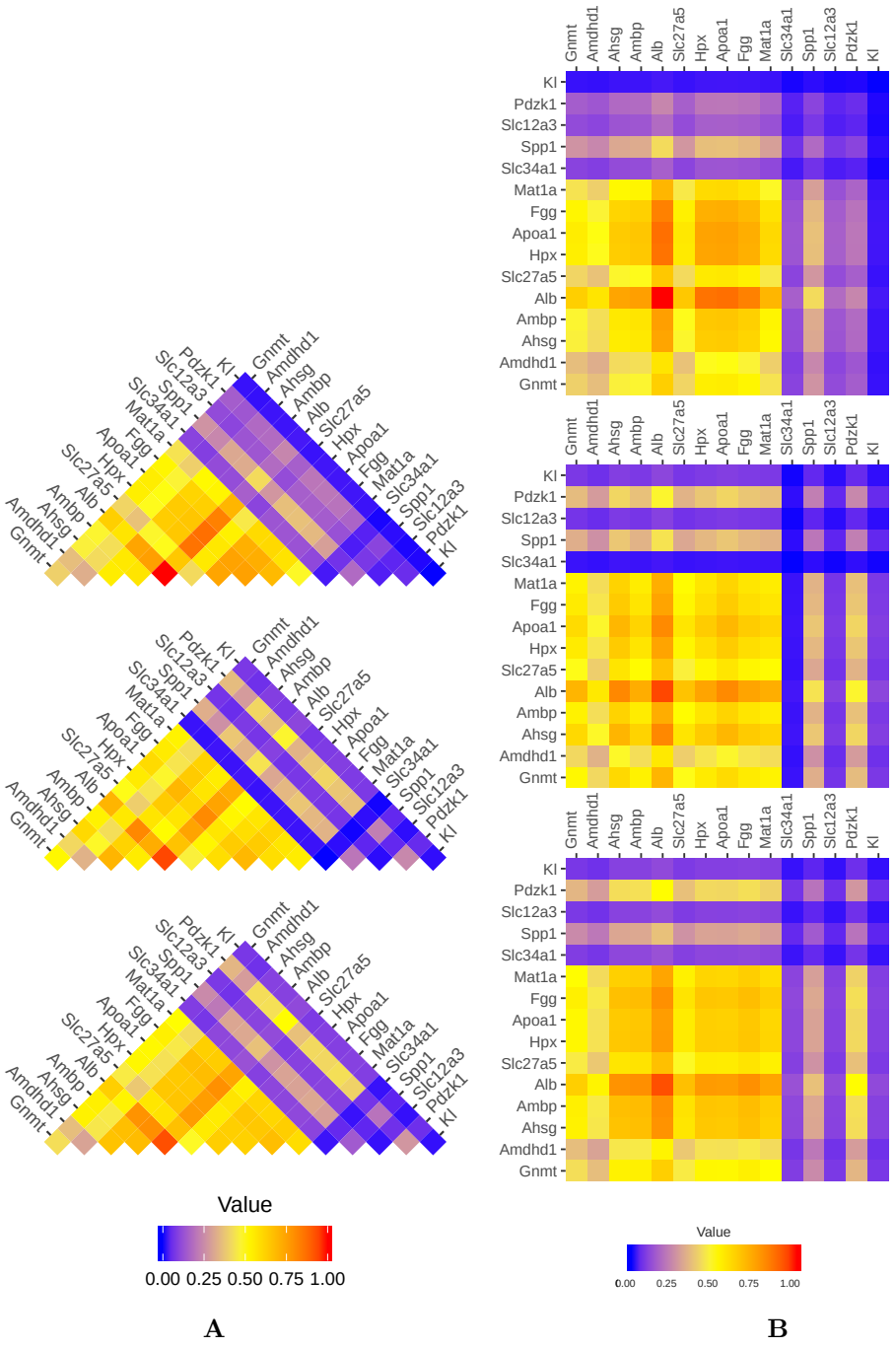


Figure A.1: Illustration of the difference between pine plots and standard heat maps. The pine plot (A) and heat maps (B) show the mutual expression of genes predicted as kidney and liver-specific in liver tissue samples across three species—macaque, chicken, and mouse. The bottom, middle, and top layers correspond to macaque, mouse, and chicken, respectively. The relationship between the expression of these genes is measured in terms of mutual expression (see Equation 3.1). The colour blue indicates that both genes are likely not expressed or expressed at very low levels. Yellow indicates that both genes could be expressed at intermediate levels or one gene could be expressed highly and the other scarcely. Red indicates that both genes are highly expressed.

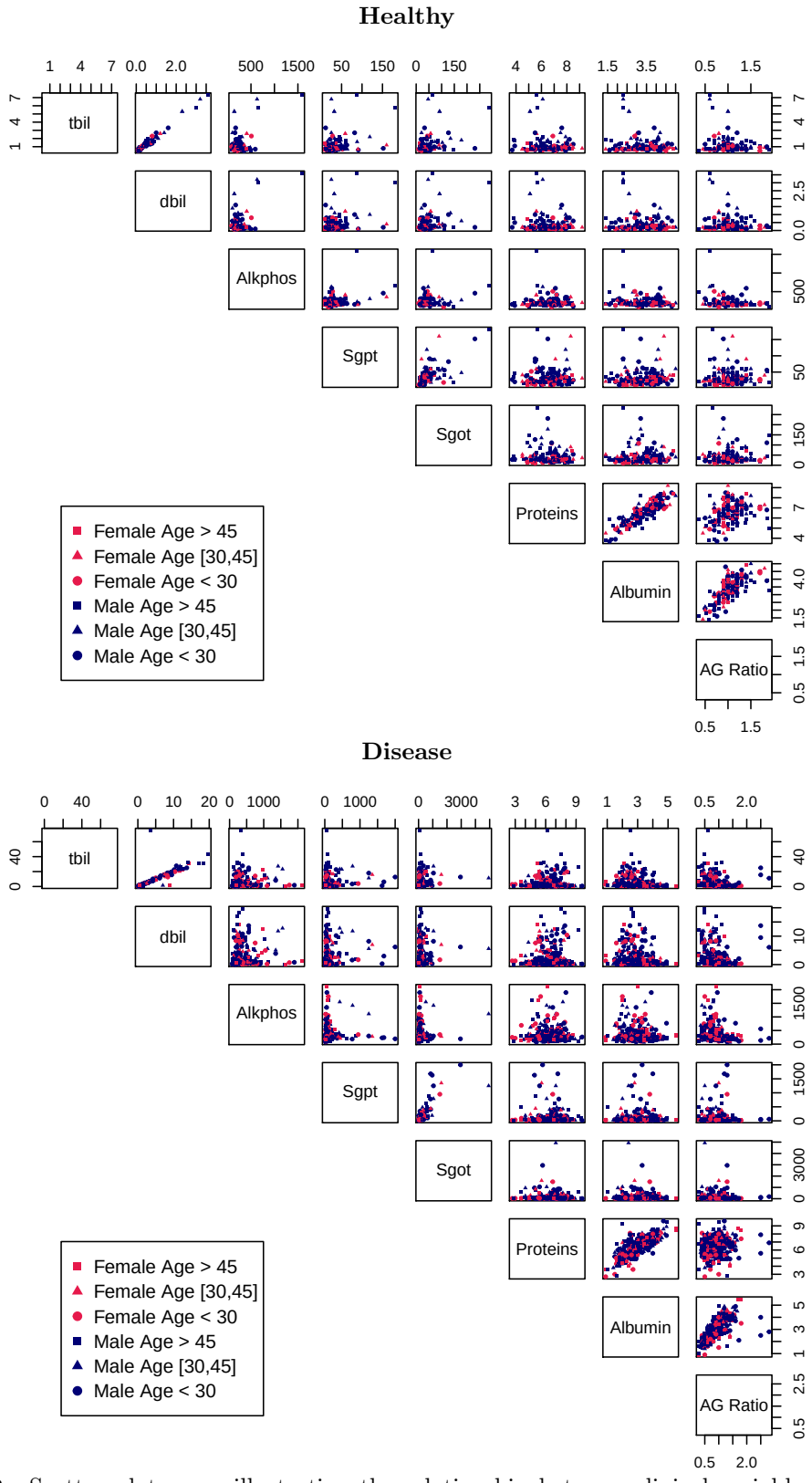


Figure A.2: Scatter plot array illustrating the relationship between clinical variables in the liver disease dataset.

Appendix B

Supplementary material for Chapter 4

Table B.1: Kruskal-Wallis Test Results Comparing Similarity Scores Across Sample Sizes

Method	Tissue	Kruskal-Wallis Test Statistic	p-value
Spearman	Brain	1884.3	<2.20e-16
	Heart	2088.9	<2.20e-16
	Muscle	2128.2	<2.20e-16
Pearson	Brain	1891.9	<2.20e-16
	Heart	2088.8	<2.20e-16
	Muscle	2120.3	<2.20e-16
WGCNA (signed)	Brain	1374.9	<2.20e-16
	Heart	1844.9	<2.20e-16
	Muscle	1860.9	<2.20e-16
Mutual Information	Brain	1801.0	<2.20e-16
	Heart	2043.7	<2.20e-16
	Muscle	2106.9	<2.20e-16

Note: A p-value of 2.20e-16 is the smallest value that can be represented using the *RVAideMemoire* R package.

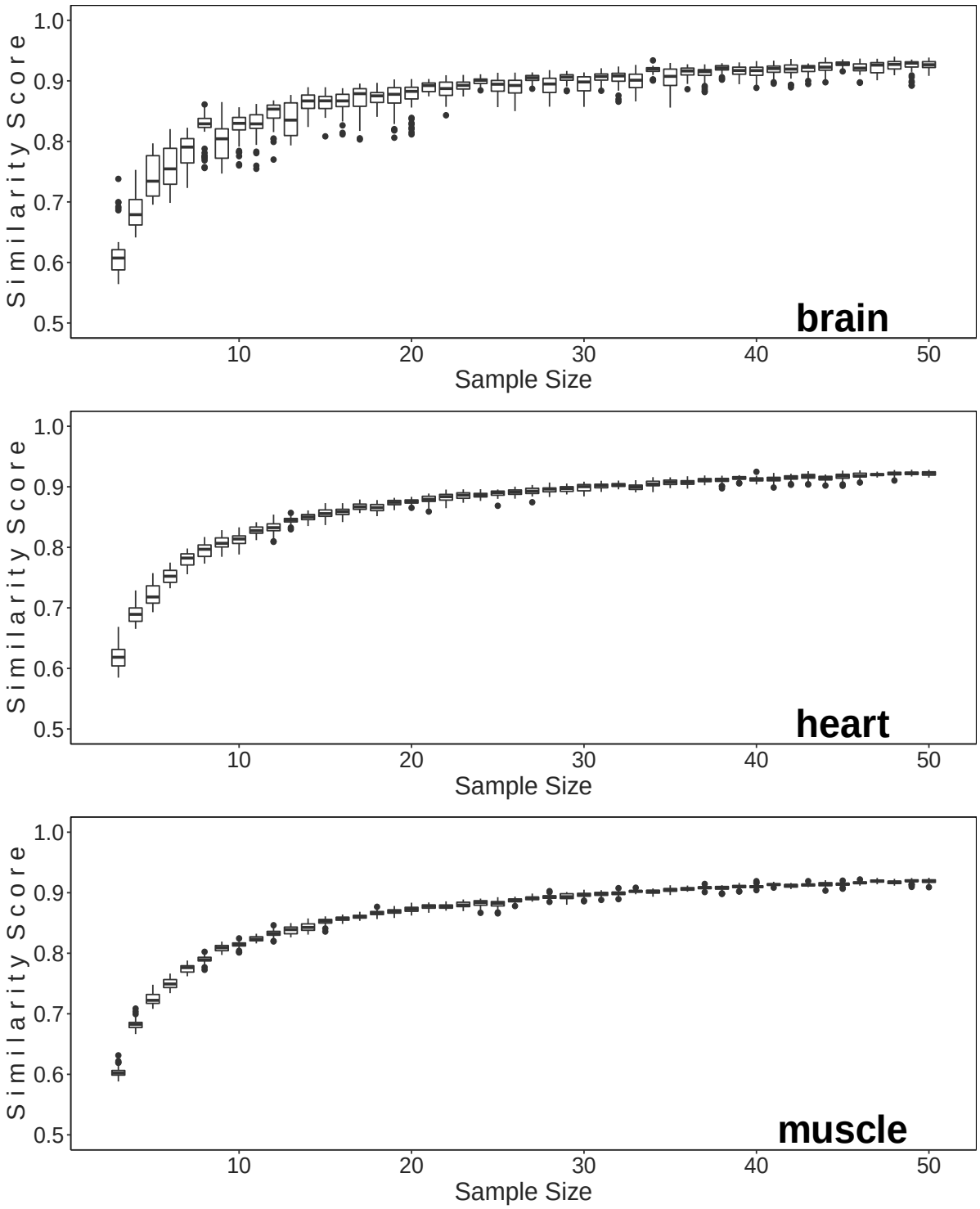


Figure B.1: Box plots illustrating the results of similarity score calculated using the normalized absolute difference between edge weights between replicate networks constructed using Pearson correlation and from 3 to 50 samples. Each sample size compares 10 replicate data sets constructed from non-overlapping samples of one of three tissues: brain (top), heart (middle), and skeletal muscle (bottom).

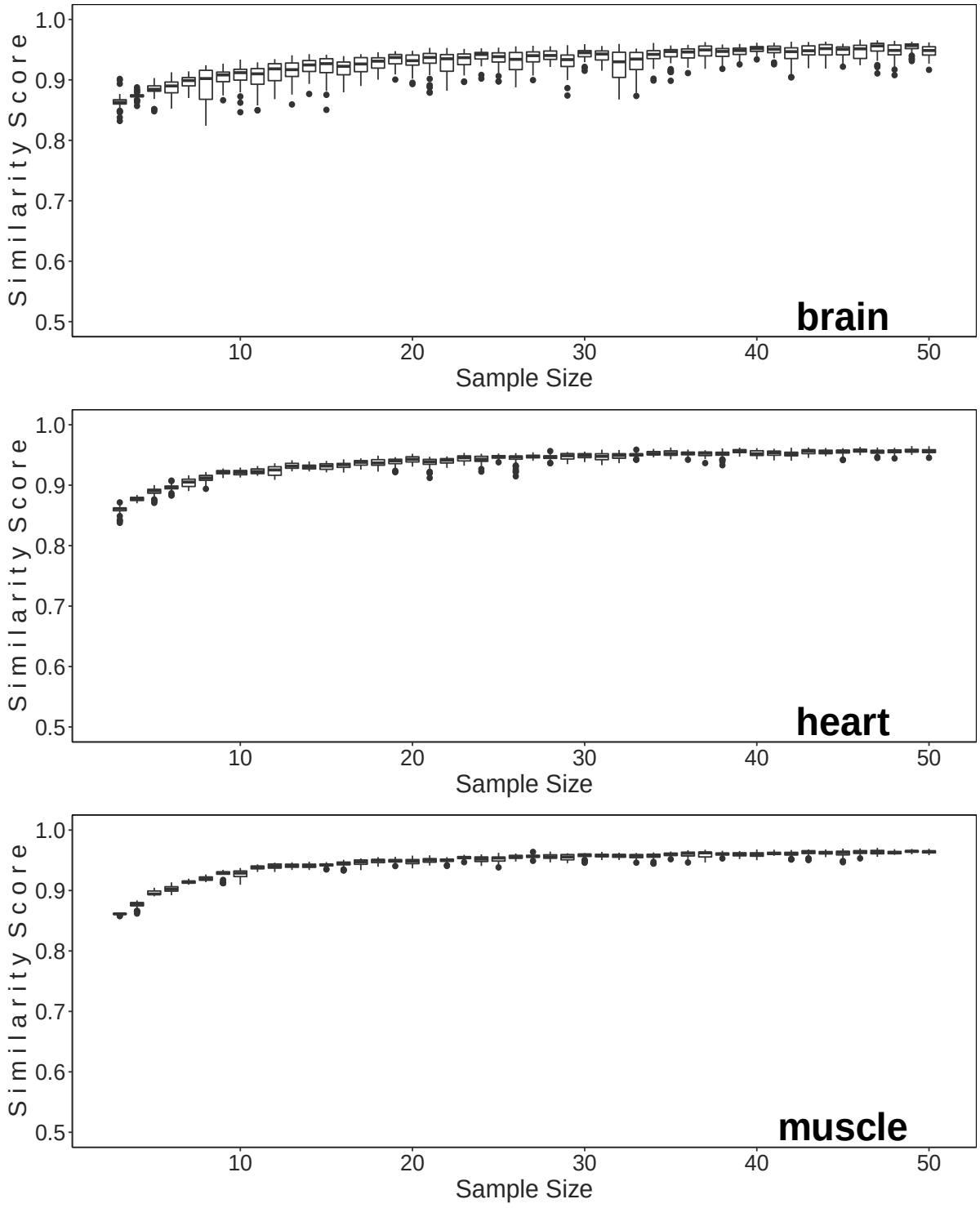


Figure B.2: Box plots illustrating the results of similarity score calculated using the normalized absolute difference between edge weights between replicate networks constructed using signed WGCNA and from 3 to 50 samples. Each sample size compares 10 replicate data sets constructed from non-overlapping samples of one of three tissues: brain (top), heart (middle), and skeletal muscle (bottom).

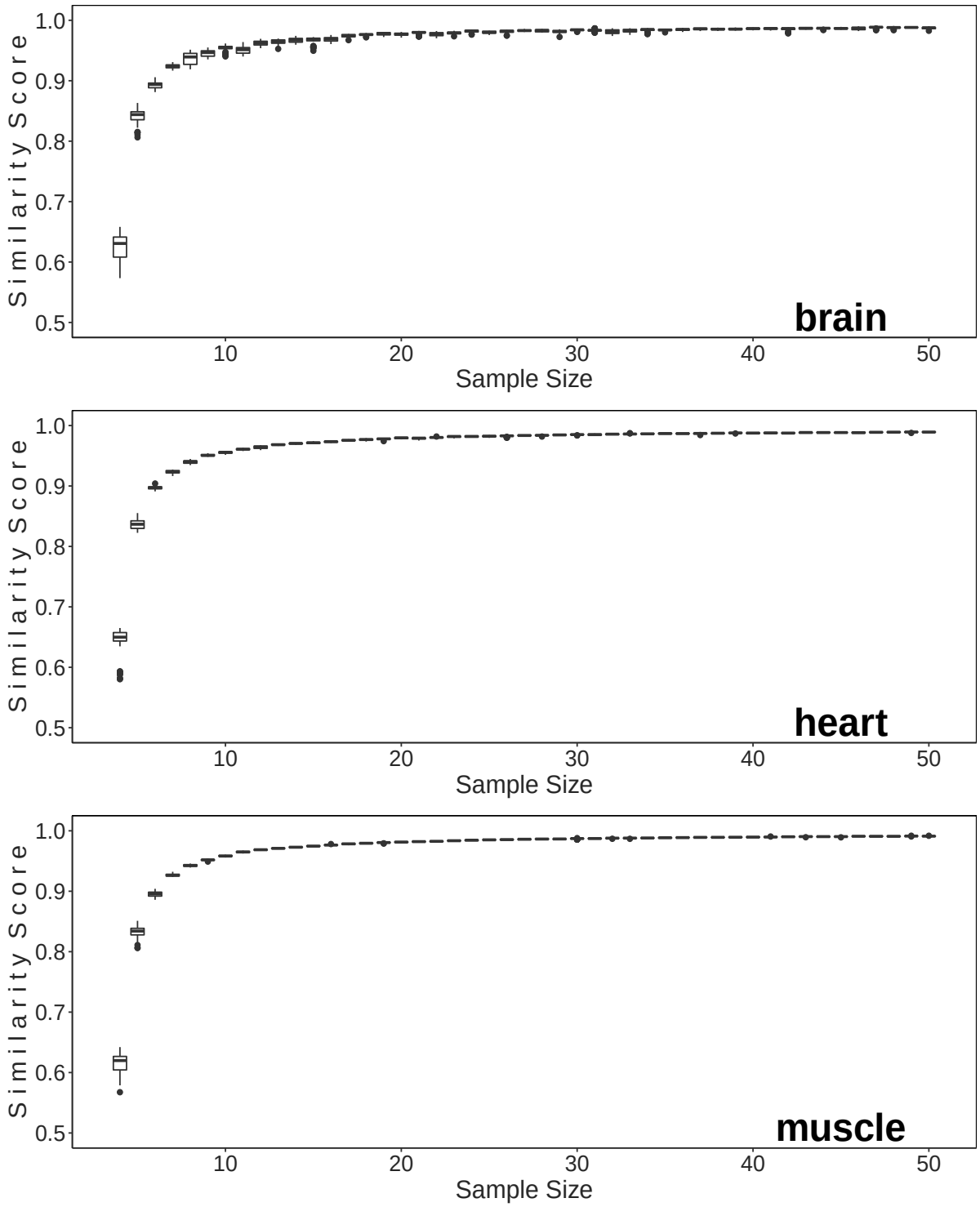


Figure B.3: Box plots illustrating the results of similarity score calculated using the normalized absolute difference between edge weights between replicate networks constructed using mutual information and from 3 to 50 samples. Each sample size compares 10 replicate data sets constructed from non-overlapping samples of one of three tissues: brain (top), heart (middle), and skeletal muscle (bottom).

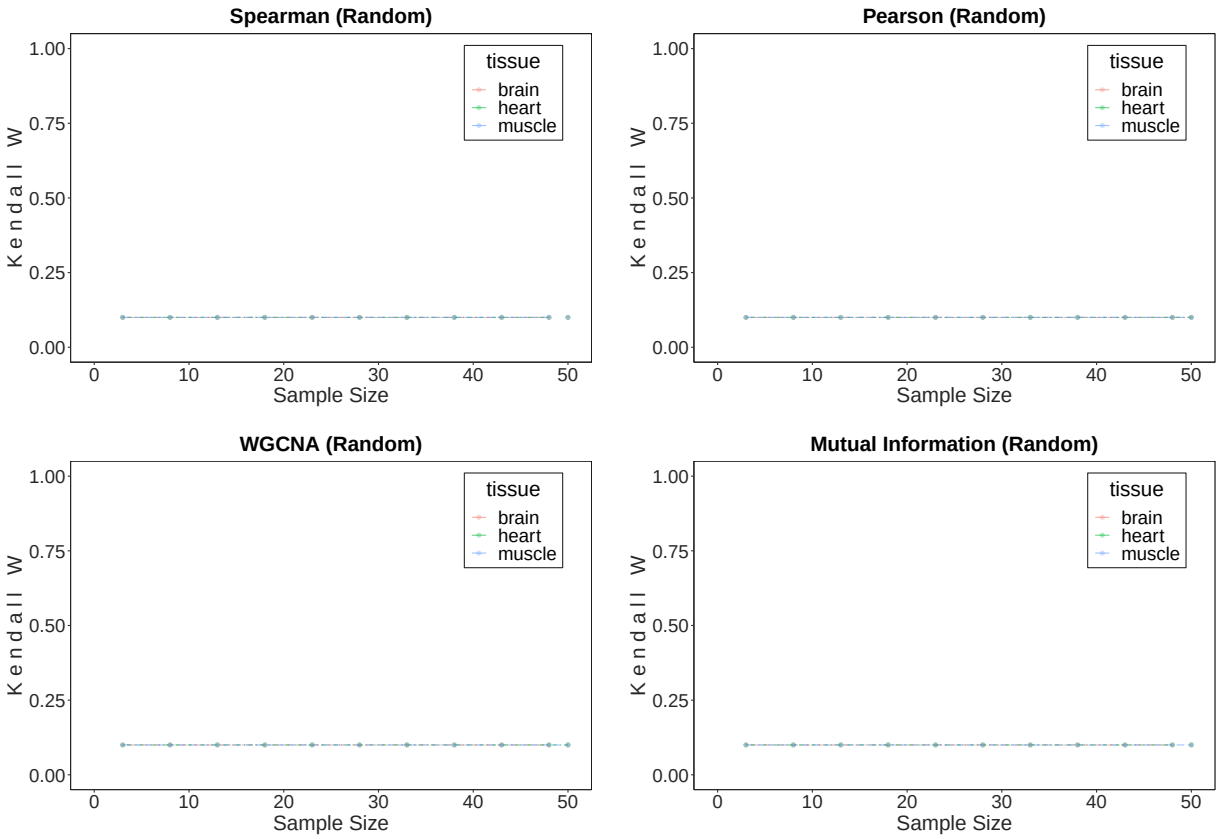


Figure B.4: Line plots illustrating the results of Kendall concordance coefficient tests for replicate networks with randomly reassigned nodes constructed from 3 to 50 samples. Each sample size compares 10 replicate data sets constructed from non-overlapping samples of one of three tissues: brain, heart, and skeletal muscle. The plots from left to right show the Kendall W values across sample sizes when constructing networks using Spearman correlation, Pearson correlation, WGCNA, and mutual information, respectively. The values of the Kendall W scores for all three tissue types are usually fluctuating between 0.08 and 0.1.

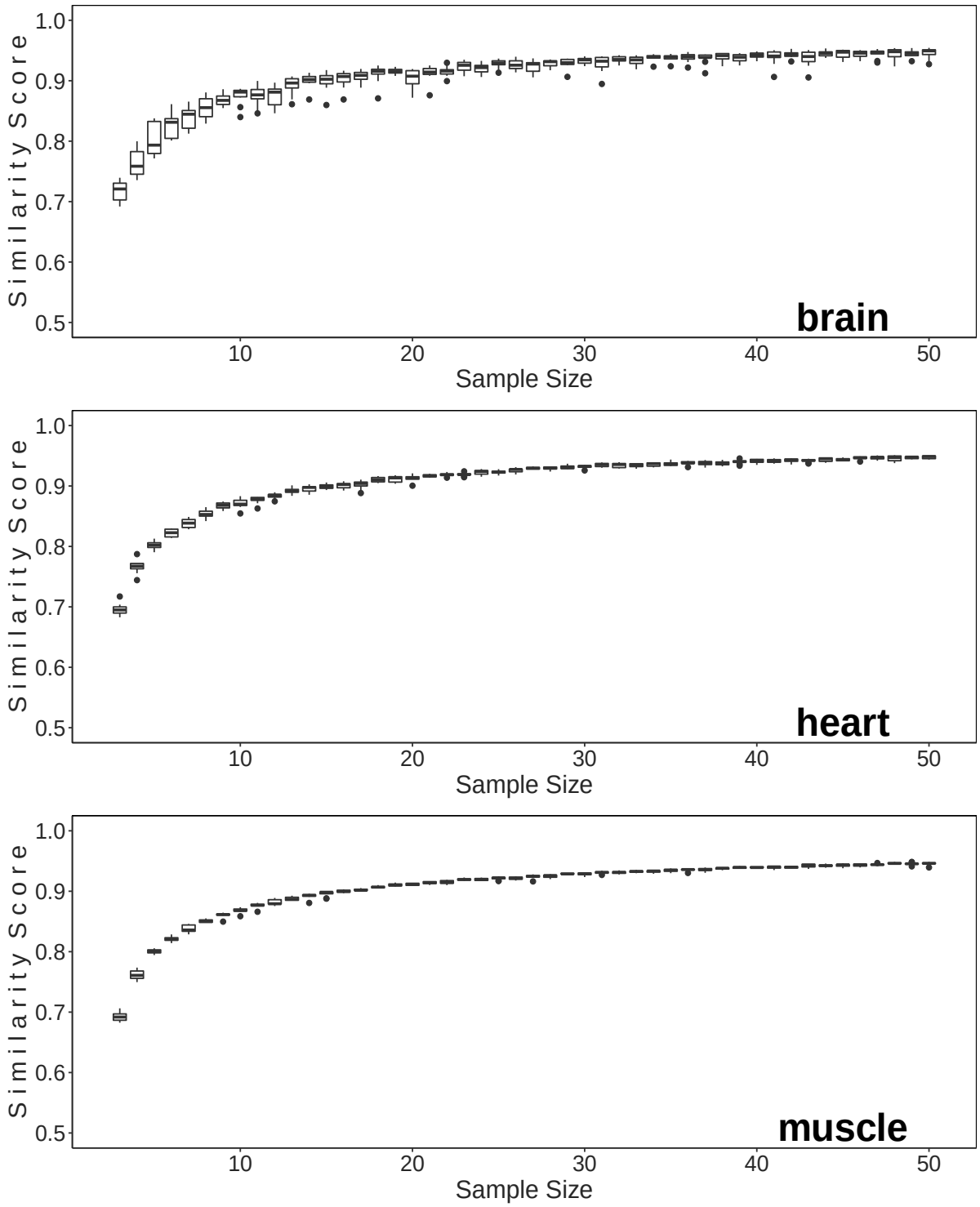


Figure B.5: Box plots illustrating the results of similarity score calculated using the normalized absolute difference between edge weights between replicate networks and the networks constructed using all available samples. These networks were constructed using Spearman correlation and from 3 to 50 samples. Each sample size compares networks constructed from one of three tissues: brain (top), heart (middle), and skeletal muscle (bottom).

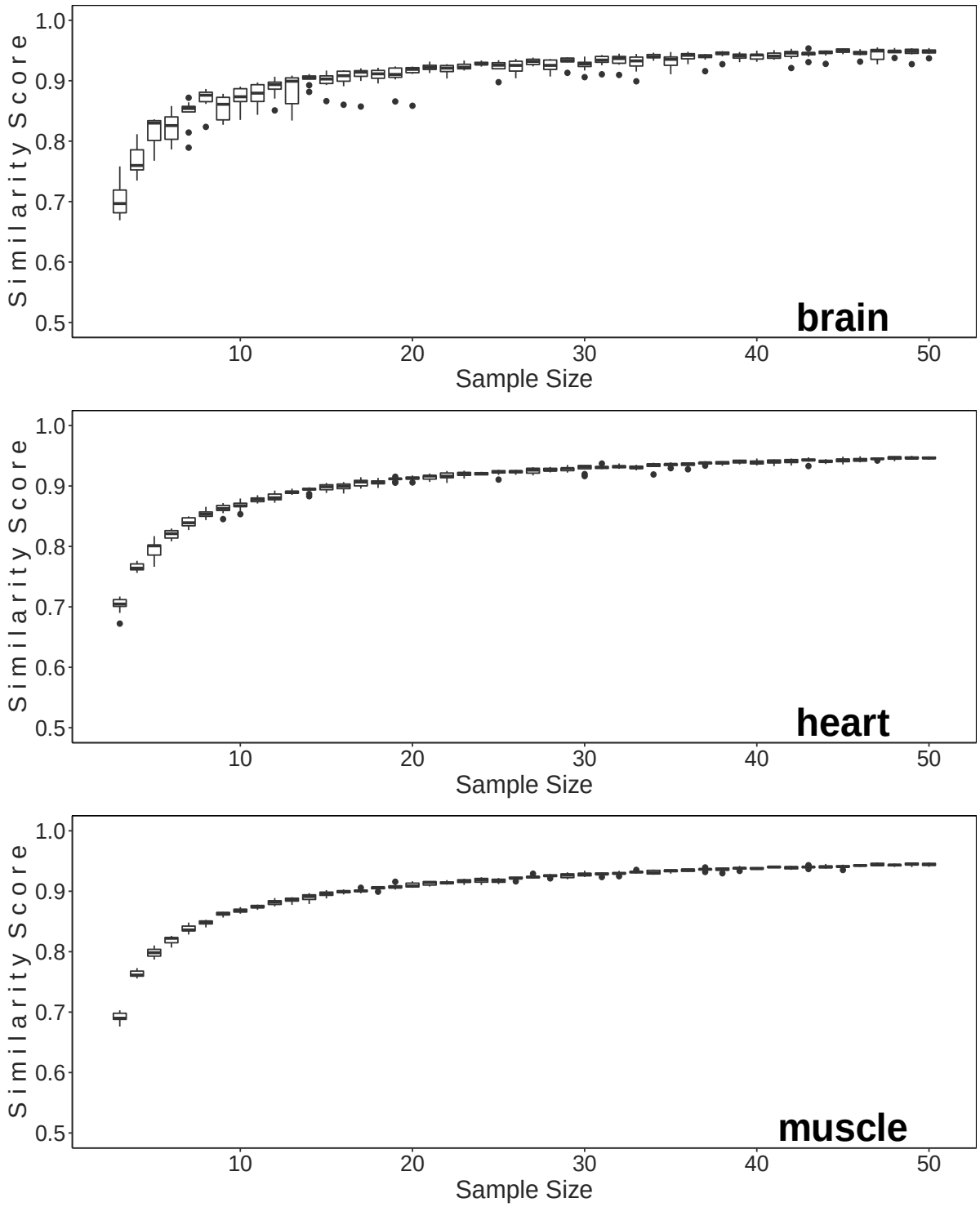


Figure B.6: Box plots illustrating the results of similarity score calculated using the normalized absolute difference between edge weights between replicate networks and the networks constructed using all available samples. These networks were constructed using Pearson correlation and from 3 to 50 samples. Each sample size compares networks constructed from one of three tissues: brain (top), heart (middle), and skeletal muscle (bottom).

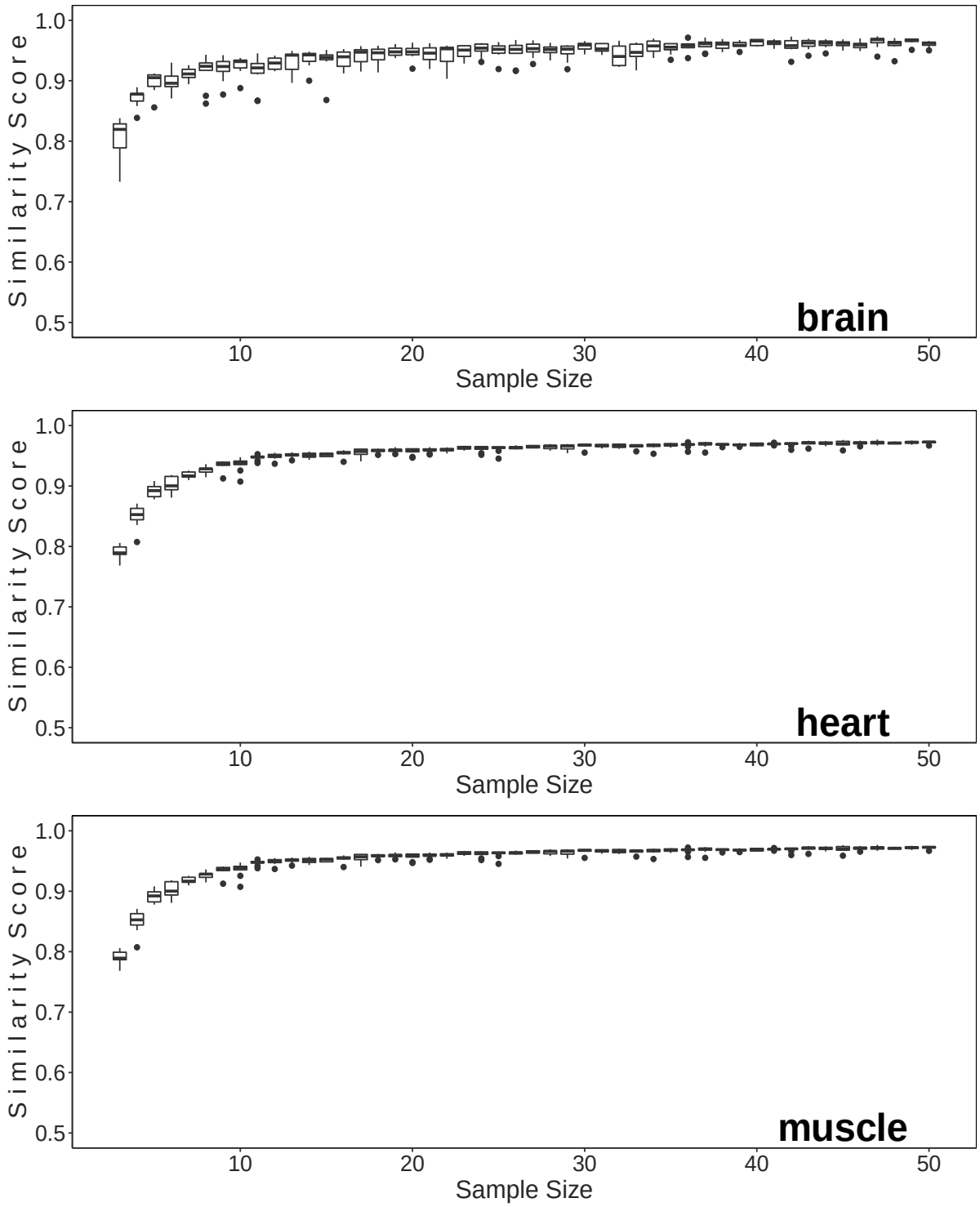


Figure B.7: Box plots illustrating the results of similarity score calculated using the normalized absolute difference between edge weights between replicate networks and the networks constructed using all available samples. These networks were constructed using signed WGCNA and from 3 to 50 samples. Each sample size compares networks constructed from one of three tissues: brain (top), heart (middle), and skeletal muscle (bottom).

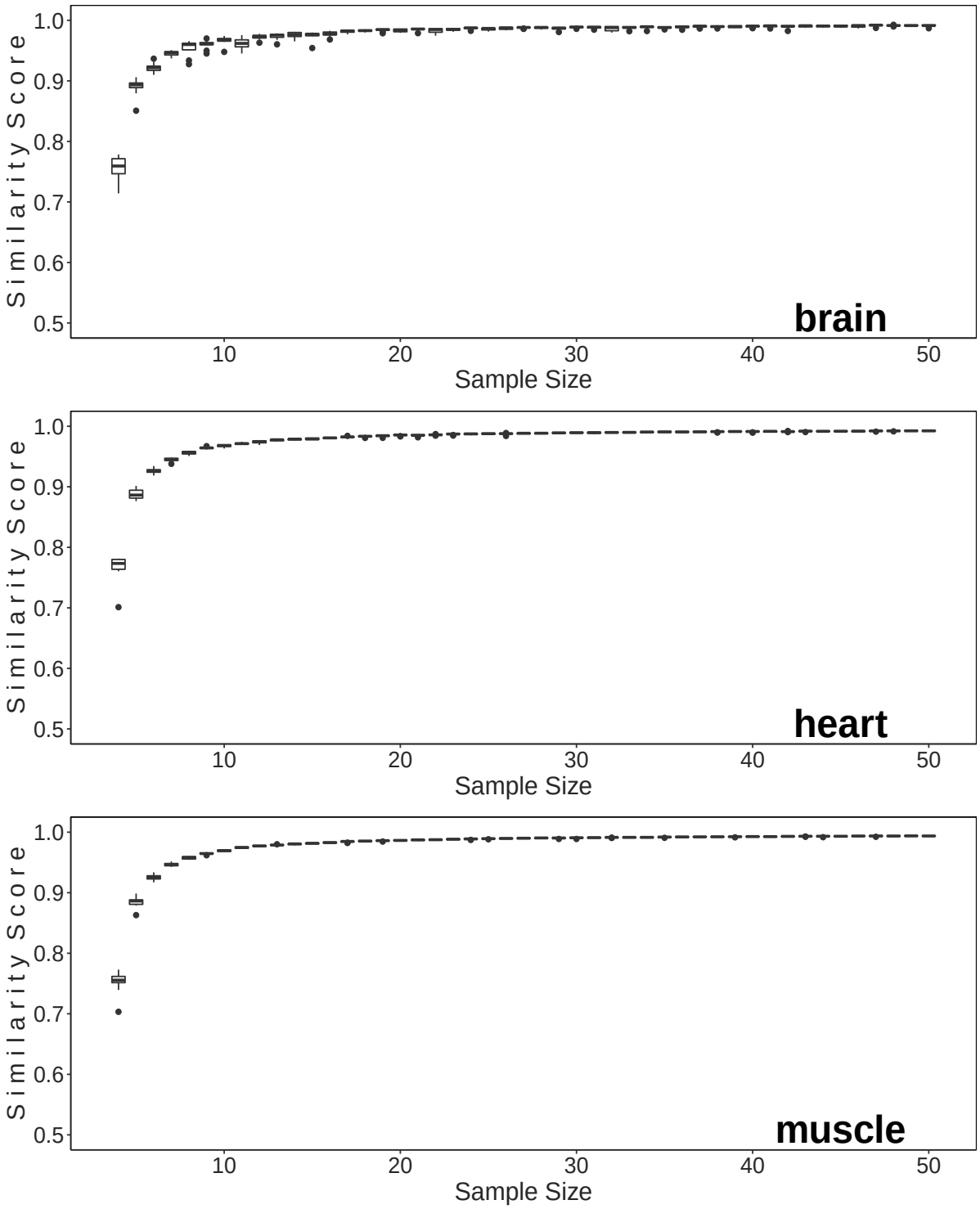


Figure B.8: Box plots illustrating the results of similarity score calculated using the normalized absolute difference between edge weights between replicate networks and the networks constructed using all available samples. These networks were constructed using mutual information and from 3 to 50 samples. Each sample size compares networks constructed from one of three tissues: brain (top), heart (middle), and skeletal muscle (bottom).

Appendix C

Supplementary material for Chapter 5

Table C.1: Parameters used for generating embedding for each GCN evaluated.

	Synthetic Networks	**Heart and Brain Networks**	**Prefrontal Cortex Networks**
walk_per_node	1000	1000	1000
walk_length	20	50	30
n_iter	1	1	1
n_workers	20	20	10
embd_dim	30	150	150
window	2	2	2
min_count	2	2	2
negatives	5	15	50
alpha	0.0025	0.0025	0.0025
min_alpha	0.001	0.001	0.001

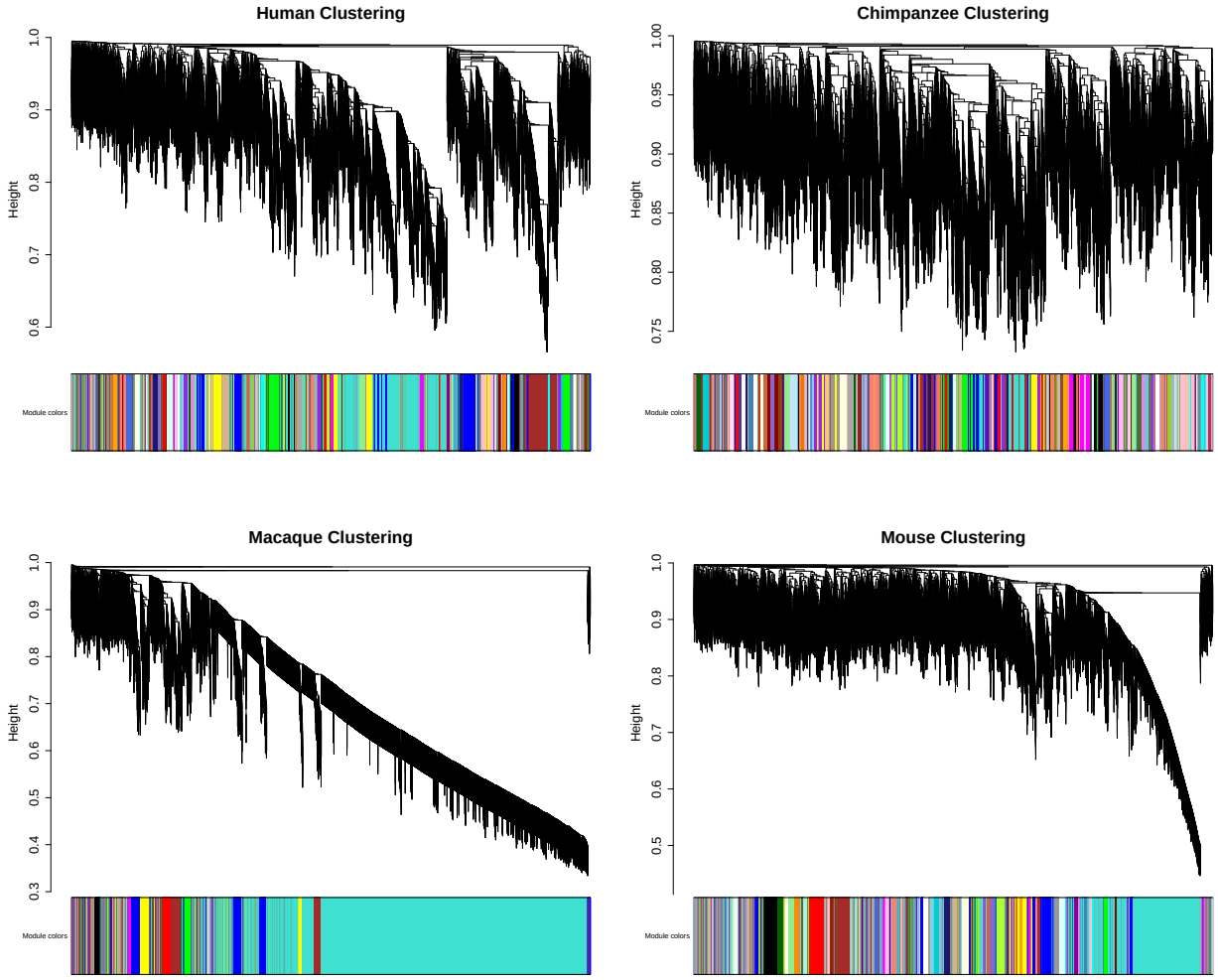
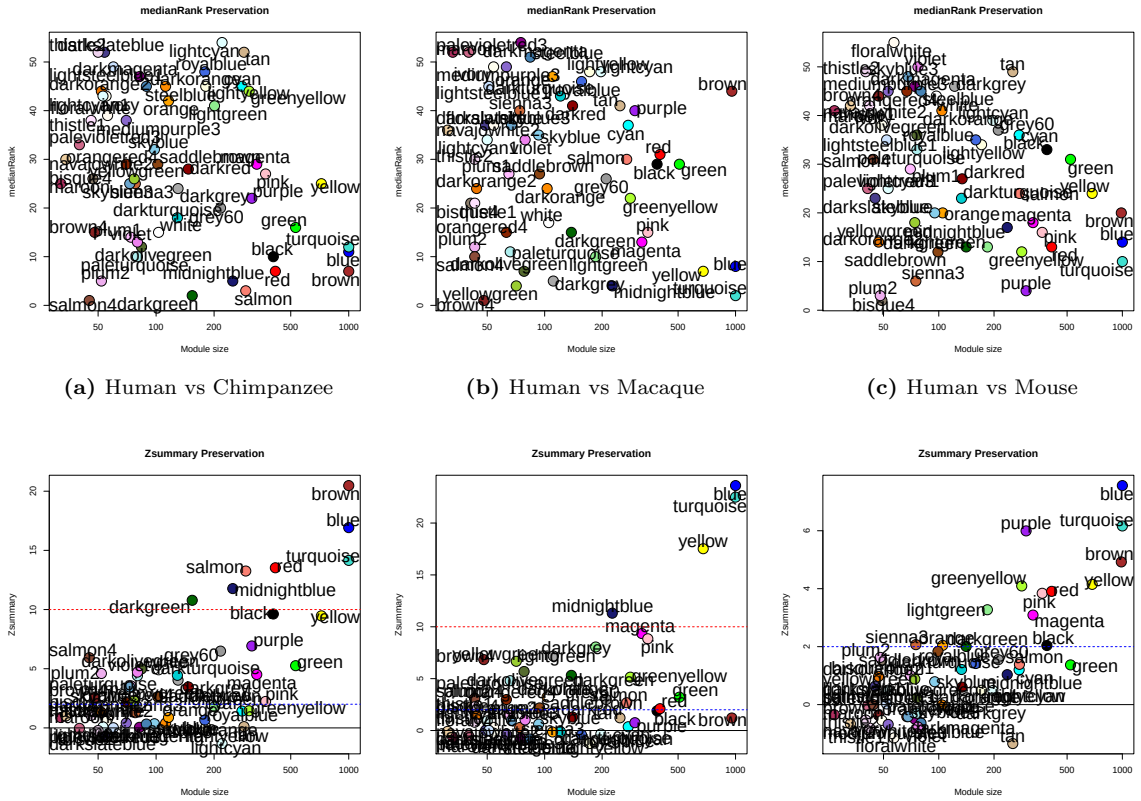


Figure C.1: Hierarchical clustering results of gene co-expression networks constructed using prefrontal cortex samples from human, chimpanzee, macaque, and mouse. To generate the networks, a power of $\beta = 8$ for soft thresholding that resulted in a scale-free network topology. The same genes used in the analysis using Juxtapose were considered for this run of WGCNA. The clustering merge height was set to 0.50 to merge to generate the clusters shown in the dendrogram and module colour images (bottom). The minimum module size allowed was 30 genes.



(a) Human vs Chimpanzee **(b)** Human vs Macaque **(c)** Human vs Mouse

(d) Human vs Chimpanzee **(e)** Human vs Macaque **(f)** Human vs Mouse

Figure C.2: Module preservation statistics comparing signed gene co-expression networks constructed using prefrontal cortex samples from human vs chimpanzee, macaque, and mouse. From both Zsummary (top left) and medianRank (top right) preservation scores using mouse as the reference network, the most preservation is observed for the yellow and brown modules. A Zsummary score greater than 2, but less than 10 indicates moderate preservation, while a score greater than 10 indicates strong module preservation. A low score for medianRank indicates high module preservation.

Appendix D

Supplementary material for Chapter 6

Table D.1: Parameters used for generating embedding for each skeletal cell GCN.

	Full Networks	**Sub-networks**
walk_per_node	1000	50
walk_length	30	10
n_iter	1	1
n_workers	12	10
embd_dim	150	5
window	2	2
min_count	2	2
negatives	50	5
alpha	0.0025	0.0025
min_alpha	0.001	0.001

Figure D.1: Heat maps of the local cosine distances showing biclustering results comparing IMM GCNs between mouse, chicken, frog, and gar. These heat maps show representative biclusters, meaning that a random selection of genes (equal to 1% of the total genes of the species) was selected in order to visualize the biclusters in a heat map. Note that enrichment analysis was performed on all the genes present in each bicluster and not just the genes visualized in the heat maps. The darker green shades indicate genes that are more similar between the networks. Light green shades indicate less similarity between the genes in a bicluster.

Figure D.2: Heat maps of the local cosine distances showing biclustering results comparing OST GCNs between mouse, chicken, frog, and gar. These heat maps show representative biclusters, meaning that a random selection of genes (equal to 1% of the total genes of the species) was selected in order to visualize the biclusters in a heat map. Note that enrichment analysis was performed on all the genes present in each bicluster and not just the genes visualized in the heat maps. The darker green shades indicate genes that are more similar between the networks. Light green shades indicate less similarity between the genes in a bicluster.

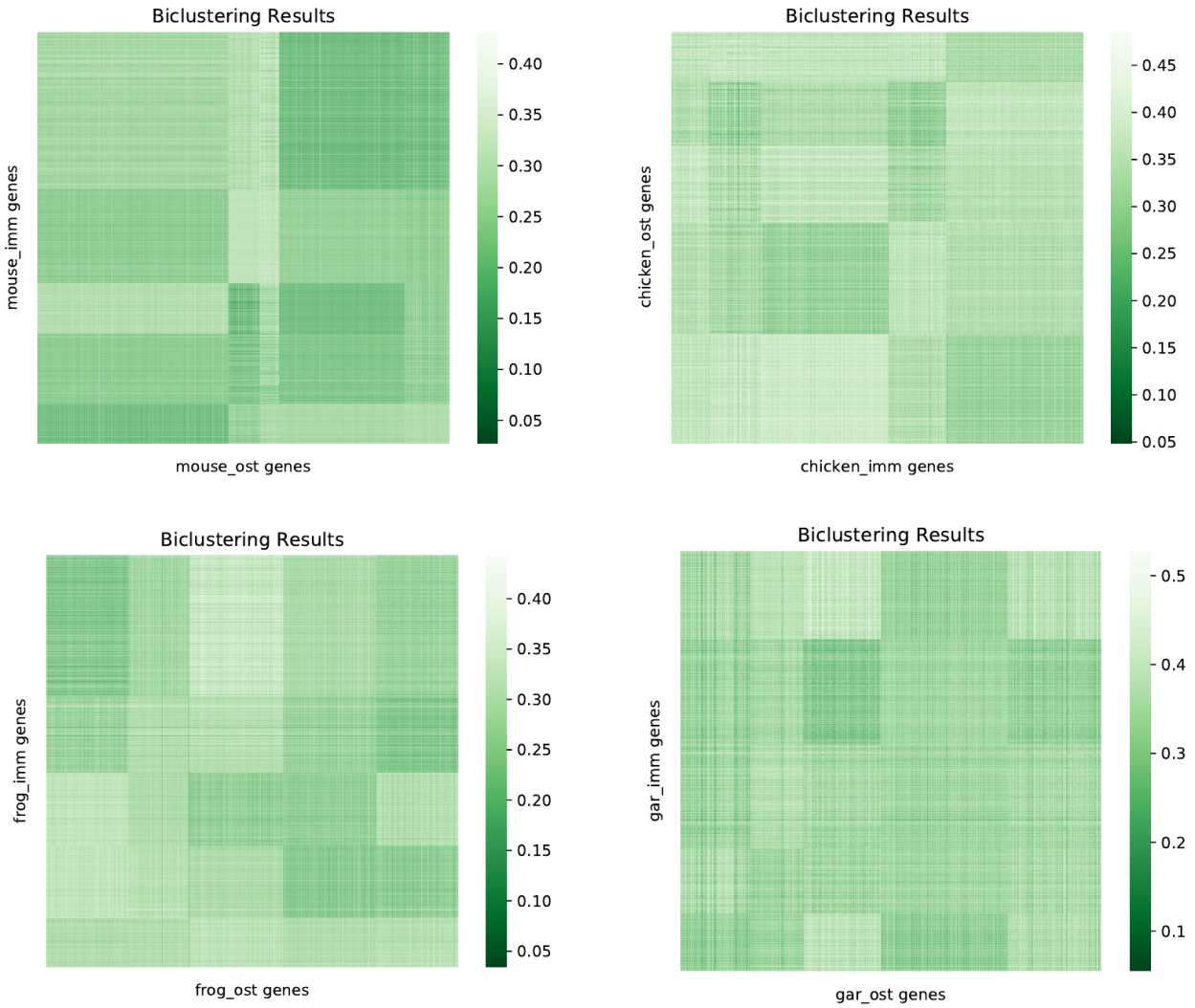


Figure D.3: Heat maps of the local cosine distances showing biclustering results comparing IMM and OST GCNs within mouse, chicken, frog, and gar. These heat maps show representative biclusters, meaning that a random selection of genes (equal to 1% of the total genes of the species) was selected in order to visualize the biclusters as a heat map. Note that enrichment analysis was performed on all the genes present in each bicluster and not just the genes visualized in the heat maps. The darker green shades indicate genes that are more similar between the networks based on their topology. Light green shades indicate less similarity between the genes in a bicluster.

Appendix E

Publication list

E.1 Peer-reviewed publications

Ovens, K., Maleki, F., Eames, B. Frank, & McQuillan, I. Juxtapose: A gene-embedding approach for comparing co-expression networks. Submitted to BMC Bioinformatics. Submission ID: BINF-D-20-01036.

Maleki, F., **Ovens, K.**, McQuillan, I., & Kusalik, A.J. (2019). Silver: Forging Almost Gold Standard Datasets for Evaluation of Gene Set Analysis Methods. Under Review at BMC Bioinformatics. Submission Number: BINF-D-19-00938.

Ovens, K., Eames, B. Frank, & McQuillan, I. (2020, June) The impact of sample size and tissue type on the reproducibility of gene co-expression networks. In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, September 21-24, 2020, Virtual Event, USA. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3388440.3412481

Maleki, F.*, **Ovens, K.***, Hogan, D., & Kusalik A. (2020, February). Gene Set Analysis: Challenges, Opportunities, and Future Research. Frontiers in Genetics 11: 654.

Ovens, K., Hogan, D., Maleki, F., McQuillan, I., & Kusalik A. (2019, December). Pine plot: Visualizing Symmetric Relationships. In Proceedings of the Tenth International Conference on Computational Systems-Biology and Bioinformatics (CSBio '19). Association for Computing Machinery, New York, NY, USA, Article 6, 1-8. https://doi.org/10.1145/3365953.3365959

Maleki, F.*, **Ovens, K.***, Hogan, D., Rezaei, E., Rosenberg, A.M., & Kusalik, A.J. (2019, July). Measuring consistency among gene set analysis methods: A systematic study. Journal of Bioinformatics and Computational Biology 17.05 (2019): 1940010.

Maleki, F., **Ovens, K.**, McQuillan, I., & Kusalik, A.J. (2019, June). Size Matters: How Sample Size Affects the Reproducibility and Specificity of Gene Set Analysis. Human Genomics 13.1 (2019): 42.

Maleki, F., **Ovens, K.**, McQuillan, I., Rezaei, E., Rosenberg, A.M., & Kusalik, A.J. (2019 September). Gene Set Databases: A Fountain of Knowledge or a Siren Call? In the 10th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB). Niagara Falls, NY, USA.

Maleki, F., **Ovens, K.**, Rezaei, E., Rosenberg, A.M., & Kusalik, A.J. (2019 February). Method choice in gene set analysis has important consequences for analysis outcome. In the 10th International Conference on Bioinformatics Models, Methods, and Algorithms. Prague, Czech Republic. 33/271 full paper submissions acceptance rate.

Maleki, F., **Ovens, K.**, McQuillan, I., & Kusalik, A. J. (2018, December). Sample Size and Reproducibility of Gene Set Analysis. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 122-129). IEEE. 105/534 full paper submissions acceptance rate.

E.2 Under preparation

Ovens, K., Eames, B. Frank, & McQuillan, I. Quantitative evaluation of evolution using comparative bioinformatics of gene co-expression networks.

Amir M. Ashique, Patsy Gómez-Picos, **Katie Ovens**, Ian McQuillan, & B. Frank Eames. GRNs driving cartilage and bone formation interact via averaging and synergism during cartilage maturation.

Amir M. Ashique, Oghenevwogaga J. Atake, Thomas Desvignes, **Katie Ovens**, Ruiyi Guo, Isaac V. Pratt, David M.L. Cooper, John H. Postlethwait, B. Frank Eames. Loss of bone trabecularity, but not bone mineral density, in Antarctic icefishes (family Channichthyidae) compared to related Antarctic notothenioids.

* First Co-author

Appendix F

Permission to reuse

Currently, the papers in this thesis that appear in literature are both published in the ACM Digital Library. Therefore, these papers may be reused in a thesis document as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included as stated on the ACM website.

Reuse

Authors can reuse any portion of their own work in a new work of their own (and no fee is expected) as long as a citation and DOI pointer to the Version of Record in the ACM Digital Library are included.

- Contributing complete papers to any edited collection of reprints for which the author is notthe editor, requires permission and usually a republication fee.
- Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included. Authors can use any portion of their own work in presentations and in the classroom (and no fee is expected).
- Commercially produced course-packs that are sold to students require permission and possibly a fee.